Understanding User Generated Content Characteristics : A Hot-Event Perspective

Guodong Li[1], Miao Wang[2], Jie Feng[2], Lisong Xu[2], Byrav Ramamurthy[2], Wei Li[1], and Xiaohong Guan[1]

[1]Xi'an Jiaotong University, Xi'an China, 710049
[2]University of Nebraska-Lincoln, NE, USA, 68588
*{lgdli,liw,xhguan}@xjtu.edu.cn, {mwang,jfeng,xu,byrav}@cse.unl.edu

***Abstract*—Nowadays, millions of Internet users watch and upload a large number of videos on User Generated Content (UGC) sites (e.g., Youtube) everyday. Moreover, online videos about hot events, such as breaking news and Olympic games, attract lots of users. In this paper, we study the characteristics of hot-event videos by collecting video traces of the largest UGC site in China for 28 days. We first empirically study statistical properties of such videos and find that hot-event videos contribute a large number of views, even though the total number of hot-event videos is relatively small. In addition, there exist extremely active uploaders and top 10% active uploaders upload over 60% videos. The video popularity demonstrates high skewness, where top 5% the most popular videos contribute over 80% views. Finally, we analyze the popularity evolution of hot-event videos using the consecutive 28-day video traces. The popularity of the studied videos decays very fast and most of these videos remain popular for only a week. Our findings reflect the most recent developments of UGC sites, which provide technical and commercial insights for engineers and UGC site owners.**

I. Introduction

Of late, Internet users are watching more and more online videos posted on User Generated Content (UGC) sites (e.g., Youtube). Comscore [1], the leading digital media measurement company, released the data of users watching online videos in the U.S. during March, 2010. According to this data, over 180 millions Internet users in the U.S. watched 31.2 billion videos in March and Youtube (the world's largest UGC site) accounted for about 41% of these videos. Moreover, UGC enables Internet users to be both video viewers and content producers, which means that the users' interests in different videos influence not only video access patterns but also video upload patterns.

The online video access patterns and upload patterns are greatly influenced by hot events, such as hot sport games and breaking news. Intuitively, users are very interested in recent hot events, in the sense that they upload and watch lots of videos about these events. According to official broadcasting statistics of Beijing Olympic Games 2008 [2], the Olympic Games generated over 628 million online videos and the video website of CCTV [3] attracted 153 million unique viewers during about one month period. Furthermore, breaking news of the death of King of Pop, Michael Jackson, generated a huge number of video uploads (e.g., memorial videos, previous music video clips etc) and views, which even crashed the Internet for a short period [4]. In this paper, we try to understand the popularity characteristics of hot-event videos by analyzing our collected latest video traces. Specifically, we answer the following questions: 1) Is hot-event UGC worth studying for designing future UGC systems? 2) What are the statistical properties of hot-event UGC? and 3) How does the popularity of hot-event UGC evolve?

To understand the UGC popularity characteristics associated with hot events, we implemented a C++ based web crawler to collect the traces of the most popular online video sites in China[1], Youku [5] (ranked 49th among all sites in the world [6] and ranked 1st UGC in China) and 6cn [7] (world ranking 738). The two sites allow users to upload contents covering a wide range of topics, which is similar to Youtube [8]. The main contribution of this paper is to analyze the statistical characteristics and popularity evolutions of UGC associated with hot events. For this, we consecutively collected video traces about hot events in China between April 12th to May 10th 2010, which account for millions of views. Based on a static snapshot of videos and views counted on May 10th 2010, our analysis reveals that hot-event UGC contributes a larger number of views with a smaller number of videos compared with data in [14], which is an earlier work of studying the Youtube video popularity in 2007. In this paper, we compare our findings with [14], which studies the world's largest UGC website Youtube and provides detailed data. Furthermore, we also find that the popularity of these hot-event UGC videos is more ephemeral than general UGC videos.

We summarize our findings as follows:

- UGC associated with hot events contribute a large number of views with a small number of videos. Compared with data in [14], the number of videos in our trace collected within 28 days accounts for only 4% of videos in [14], which collected videos uploaded in more than a year. However, the number of views accounts for about 25% of views in [14]. It indicates that studying characteristics of such videos is helpful in UGC system design.
- Uploading skewness and popularity skewness are important characteristics of UGC related to hot events. First,

[1] At the time when we developed and deployed the crawler, we encountered the access problem to Youtube from China. We believe that the findings in this paper should have great similarities to Youtube, because all these sites are all general purpose UGC sites. We will revise the parser and deploy our crawler from a server outside of China to crawl Youtube soon.

top 10% uploaders upload over 65% videos and there exist extremely active uploaders, who upload more than 1,000 videos within 28 days; whereas heavy uploaders discovered in [14] take a few years to upload 1,000 videos. Second, the views of hot-event videos are highly skewed and top 5% videos contribute over 80% views, which indicates that viewers' interests are biased towards some hot videos.

- The popularity of hot-event UGC decays very fast and most videos remain popular only for a short time (about 7 days). 90% of videos reach their view peaks within 7 days after uploaded. Based on a static snapshot on May 10th 2010, the popularity distribution of hot-event videos demonstrates a Zipf-like waist with a dropped tail, which is similar to the results of general UGC in [14].

The rest of the paper is organized as follows. We summarize related work in Section II. Section III introduces our collection methodologies and data sets. In Section IV, we analyze our data sets and describe the results. Finally, we conclude our paper in Section V.

II. Related Work

In term of UGC popularity research, there are two works closely related to this paper. Cha *et al.* [14] analyze popularity characteristics of UGC systems using traces collected from Youtube and another UGC website in Korea. Our work differs from [14] as follows: 1) We try to understand the UGC popularity characteristics associated with hot events, which can be considered as a micro analysis of popularity; whereas [14] provides a global view of popularity; 2) The data used in [14] were collected about one year after Youtube first being online in 2005, which corresponds to early days of UGC sites; and 3) Hot events are usually ephemeral and thus we can analyze the popularity in a full life cycle. To sum up, our analysis can provide insights into short term content placement and dealing with user upload and access bursts. By contrast, [14] can help with long-term content placement. Cheng *et al.* closely look at the social networking aspect of Youtube and conclude that the videos have strong correlations with each other, which can be used to improve the streaming quality [15].

Content popularity is essentially important in design and implementation of video streaming systems and there are a few studies of content popularity in systems other than UGC. Qiu *et al.* [18] focus on analyzing the channel popularity of Internet Protocol Television (IPTV) systems and establish a stochastic model to capture the characteristics of channel popularity. However, IPTV systems are different from UGC sites, in the sense that contents in IPTV systems are generated by content providers, which is similar to traditional cable TV systems and Video-on-Demand (VoD) systems. Similarly, Allen *et al.* [12] study the popularity distribution in cable networks. Huang *et al.* [17] study the feasibility of using Peer-to-Peer (P2P) streaming technologies in online video sites, via analyzing MSN video server logs. In addition, finance research shows that breaking news have great impacts on stock returns [13] [16]. UGC popularity might also be greatly influenced by hot-events, which motivates our measurement study in this paper.

III. Data Collection

In this section, we introduce our data collection methodology and provide a high-level overview of our data set.

A. Measurement Methodology

We collect data traces of two UGC sites, Youku and 6cn, where Youku ranks 1st and 6cn ranks 9th among all UGC sites in China [6]. The two sites are very similar to Youtube and attract millions of users everyday. There is no login requirement for watching videos, but a user is required to login for uploading, commenting and rating videos, which is the same as Youtube. Both Youku and 6cn stream videos via the Adobe Flash Player plugin embedded in user's web browser. In addition, there is no video length limit in both sites, although they encourage users to cut large video files into smaller pieces with provided tools. Compared with Youku and 6cn, Youtube imposes the 15-minute video length limit [8].

To understand UGC popularity associated with hot events, we take advantage of video search engines provided by the two sites, which take search key words as input and return pages containing related videos. We first collect hot events occurring between April 12th and May 10th, based on hot topics posted on portal sites (e.g., SINA [9]) and popular forums (e.g., MOP [10]). People are highly interested in these events, since they range from breaking news (e.g., earthquakes, coal mine disasters) to famous social events (e.g., Expo 2010 Shanghai China) during the observation period. Next, we record the key words for corresponding events into a configuration file. Note that the events are usually added into the file immediately after we see the related hot topics, which might not capture the immediate occurrences of some events. However, the evolution trend of videos related to the events should not be greatly influenced by the small delay. Our C++ based crawler reads the file and sends HTTP requests containing these key words to search engines. Finally, our crawler parses the returned HTML pages containing video information (e.g., video length, video url, uploader, viewed times and comments etc.) and stores the classified records into files based on different events. In order to capture the popularity dynamics, we crawl the two sites 6 times a day and consecutively crawl for 28 days.

B. Data Set Overview

We summarize our data sets in Table I, which provides a high-level description of data traces with basic statistics. The statistics are calculated on May 10th, which include all captured *unique* videos during the observation period (from April 12th to May 10th). As shown in Table I, both Youku and 6cn have two kinds of data sets: 1) video traces that include all videos corresponding to all hot events (i.e., Youku-All and 6cn-All); and 2) video traces that exclude videos corresponding to hot TV episodes during the observation period (i.e., Youku-No-TV and 6cn-No-TV). Users can upload full length TV episodes in both Youku and 6cn, which greatly

TABLE I
SUMMARY OF DATA SETS

Name	Num. events	Num. unique videos	Total views	Total length	Data collection period
Youku-All	62	110,423	1,517,292,947	3.4 years	April 12-May 10 (6 crawls/day)
Youku-No-TV	51	72,785	921,516,791	1.5 years	April 12-May 10 (6 crawls/day)
6cn-All	56	83,140	522,383,251	1.77 years	April 21-May 10 (6 crawls/day)
6cn-No-TV	47	79,812	506,780,329	1.73 years	April 21-May 10 (6 crawls/day)

TABLE II
DATA SUMMARY IN CHA07

Name	Num.videos	Total views	Total length
Youtube-Ent	1,687,506	3,708,600,000	15.2 years
Youtube-Sci	252,255	539,868,316	1.8 years

influences the length distribution of observed videos, because there is no video length limit in the two sites.

Before discussing our findings in the following section, we first estimate the contributions of hot-event UGC in terms of the number of views. Since we do not crawl all videos in Youku and 6cn, we use the data set about the number of videos and views in Youtube [14] (referred to as Cha07) for our estimation, which is shown in Table II. Note that our estimation is conservative, because Youtube is known as the world's largest UGC site. Based on data set in Table II, we have the following findings: 1) The total number of videos in our traces are much smaller than that in Cha07, which indicates that the number of hot-event videos in a month period accounts for a small fraction of all videos; and 2) hot-event videos contribute a large fraction of views. For example, the average number of views per video of Youtube-Ent and Youku-All are 2,197 and 13,741 respectively. The basic statistics imply that studying characteristics of hot-event videos are important for future UGC systems design.

IV. CHARACTERISTICS OF UGC OF HOT EVENTS

In this section, we introduce our findings of UGC characteristics related to hot events. We use methods similar to [14] to compare our findings with general UGC characteristics. Then, we describe specific popularity characteristics of UGC associated with hot events.

A. Empirical Statistic Properties

In this subsection, we describe the production speed of contents associated with hot events, the high popularity skewness, the video length disribution and the quantity distribution over different events. The following observations are based on the snapshot on May 10th, which is the last day of our data collection and includes all unique videos collected between April 12th to May 10th.

1) How do heavy producers contribute to uploads?: As shown in [14], UGC requires less production effort and accordingly has a larger number of distinct content producers. Therefore, it is not very surprising that 26,413 unique producers uploaded 110,423 videos (4 videos per user) during observation period based on Youku-All data set. However, it is surprising that the major proportion of these videos are uploaded by a small group of producers and there exist some extremely active producers, who uploaded over 1,000 videos within less than one month. In contrast for the data set in [14], the heavy producers upload about 1,000 videos over *a few years*. To illustrate the above finding, we count the number of uploads per producer and show it in Figure 1. The X-axis denotes the producers sorted from the most active to the least active, with ranking of upload numbers normalized from 0 to 100. This figure shows that the top 10% producers account for 65% uploads and the top 30% producers account for 80% uploads.

2) High popularity skewness: The popularity of UGC related to hot events shows very high skewness, which can be seen from Figure 2. We rank all videos in the Youku-All trace based on the number of views, from the most popular to the least popular videos. Then, we normalize video ranking from 0 to 100 and arrange them along the X-axis. According to Figure 2, top 5% videos contribute over 80% views, which demonstrate higher skewness than general UGC videos [14] (top 10% videos contribute about 80% views). We also plot video ranking against views in log-log scale in Figure 3, where unpopular videos occur on the the tail. As shown Figure 3, the UGC associated with hot events demonstrates a Zipf-like waist with a dropped tail.

The above findings show that users always focus on a very small number of videos about the same hot event. Based on our observation, there are two possible reasons for the higher skewness: 1) The video search engine provided by Youku returns almost the same first page of search results when searching at different times and users might only watch the first few videos on the first page; and 2) Youku usually recommends videos with high numbers of views to the homepage, which leads to the *rich-get-richer* result.

High skewness indicates both business and administration opportunities. Since top 5% videos about hot events attract over 80% of views, video sites can hold auctions for posting advertisements on these pages along with the occurrence of hot events, which help advertisers spend their money on pages with maximum exposures. In addition, video sites can strategically add more interesting video links to attract more clicks on other pages. Site administrators can also take the advantage of the skewness in balancing server load (e.g., caching most viewed videos evenly among different servers).

3) Video length: Figure 4 shows the video length distribution of Youku-All and 6cn-All data traces. Compared with Youtube videos, about 40% Youku videos and 20% 6cn videos are longer than 20 minutes, because there is no video length

Fig. 1. Skewness of upload quantities across users

Fig. 2. Skewness of user interests across videos

Fig. 3. Youku-All empirical plot in log-log scale: Zipf-like waist with a dropped tail

limit in both sites. The median video lengths for Youku and 6cn are 373 seconds and 240 seconds, respectively. Compared with Youku, 92% videos in 6cn are shorter than 30 minutes and there are almost no videos with lengths between 30 minutes and 84 minutes. It might be because 6cn is less popular than Youku and users are less likely to upload full TV episodes to 6cn. However, there are a small number of videos (4%) in 6cn with lengths over 90 minutes.

4) Distributions over different events: As mentioned in Section III-B, we observed 61 hot events in Youku-All data set between April 12th to May 10th 2010. We list some of events as examples: 1) breaking news: Wangjialing Coal Mine Disaster, in which 200 people died; 2) hot movie: Alice in Wonderland; 3) hot TV episodes: Hi, My Sweetheart; and 4) hot event: Expo 2010 Shanghai China. In this subsection, we show the distributions of different attributes over the 61 events. In all the following figures, we rank the 61 events along the X-axis based on the number of videos corresponding to each event, from the largest number to the smallest number. Figure 5 shows the distribution of comments over different events, which does not strictly decrease with the decrease of the number of videos. The largest increase of comments occurs at the event with the 10th largest number of videos, which corresponds to the earthquake in mid-April. It indicates that breaking news are very likely to increase users' participation. However, hot events do not necessarily imply high user participation. In Youku-All data set, there are 2,438,894 comments of all events; whereas there are 1,517,292,947 views of all events. The ratio of comments divided by views is only 0.16%, which is exactly the same as the ratio provided by [14]. Since UGC sites are usually considered as an example of Web 2.0 sites, this result also matches the participation inequality in Web 2.0 sites [11] (i.e., only very small number users participate in uploading and commenting posts). One implication is that the comments on videos might not be representative samples for marketing research or estimating video popularity.

We show the number of views over 61 events in Figure 6. Compared with Figure 5, this figure also demonstrates the trend that the number of views does not strictly decrease with the decrease of the number of videos. Moreover, the event with the 3rd largest number of videos in Figure 6 has almost the lowest number of views among the top 10 events. The event is about the movie Alice in Wonderland, which was first shown in China mainland on March 26th 2010. We manually search these videos with the same key words used in our configuration file for that event. We find that all the videos are about reviews of the movie and short clips (within 1.5 minutes). It seems that users are interested in full movies instead of reviews. The event ranked 13th in Figure 6 generates a peak of views, which is corresponding to a hot Chinese TV show (Hi, My Sweetheart). We manually search the videos and find that there are many full length episodes contributing a large volume of views.

B. Popularity dynamic characteristics of UGC with hot events

In Section IV-A, we study the global views of UGC associated with hot events. In this subsection, we focus on popularity evolutions along with corresponding hot events based on consecutive day observations. First, we study the probability of videos being popular over time (on a daily basis). Second, we define the Video Active Interval (VAI) as an index to study how long the videos remain popular. Finally, we study how fast the videos reach the peak of view increment.

1) Probability of videos being popular over time: The popularity of videos decays with the increase of age. We define the $\Delta views$ of a video as follows: for a video A, given two consecutive days N and $N-1$ with corresponding numbers of views V_N and V_{N-1}, then $\Delta views$ of A between N and $N-1$ is $V_N - V_{N-1}$. We only consider the videos, which are observed on April 12th and exist during the whole observation period (64,671). Figure 7 shows the percentage of videos with $\Delta views \leq V$ over the observation period. We change V from 5 to 10,000. From Figure 7, we can see that from day 1 to day 3, the percentage deceases, which indicates that the popularity of hot-event videos increases after these events occurring a few days. After day 4, the percentage increases against time for all V values, which means that the popularity begins to decay. After day 7, when $V > 100$, the percentage reaches about 90%, which means that users are not interested in 90% of videos about hot events that occurred one week ago.

2) Active time distribution: In this subsection, we study how long videos associated with hot events remain popular. For this, we first define the Video Active Interval (VAI) to measure the active days of a video.

Video Active Interval (VAI): With a defined view increment threshold V and a given video A, we search view increments (i.e., $\Delta views$) of A during the observation period and find out the following two days: the first day, D_{first} that $\Delta views$ of A are greater than or equal to V; and the last day, D_{last} that $\Delta views$ of A are greater than or equal to V. The VAI of A is defined as $D_{last} - D_{first} + 1$.

In Figure 8, we show the VAI of videos in Youku-All data set and provide several view points by considering V values with a range from 5 to 10,000. When $V = 5$ (i.e., once there are more than 5 views of a video on a specific day, we consider

Fig. 4. Video length distributions of Youku-All and 6cn-All

Fig. 5. The comments of videos associated with 61 hot events of Youku-All data set

Fig. 6. The views of videos associated with 61 hot events of Youku-All data set

Fig. 7. The probability of videos with $\Delta views \leq V$ increases over time with Youku-All data set

Fig. 8. CDF of Video Active Interval with Youku-All data set

Fig. 9. The CDF of reaching view peak, 90% of videos reach peak within 7 days

the video is active on that day), about 40% videos have 1-day active intervals. When $V \geq 100$, about 90% videos have 7-day active intervals. Moreover, when $V \geq 100$, the lines between 7 and 16 are relatively flat, which indicates that there are few videos with active intervals between 7 and 16. There are 4% to 5% videos active during the whole 28-day observation period. It confirms the result in Section IV-A2 that top 5% videos contribute over 80% views.

3) View increment speed for newly uploaded videos: Since we consecutively collect video traces with occurrences of corresponding hot events, it is possible for us to study the popularity evolution starting from when a video is first uploaded. To achieve our goal, we first search daily data traces to filter out the videos with 0 view and mark their initial occurrence days as corresponding birthdays. We find 15,301 such videos, because we crawl data in fixed times, but users continually upload and watch videos. Moreover, we find that 12,427 of the 15,301 videos only exist 1 or 2 days, which is mainly due to the following two reasons. 1) These videos violate posting rules and are deleted very soon; and 2) Video search engine provided by Youku filters out some videos. Therefore, we finally obtain 2,874 videos existing more than 2 days, whose birthdays fall within our observation period.

For the 2,874 videos, we first calculate their respective daily $\Delta views$. Then, for each video, we calculate the number of days elapsed between its birthday and the day reaching the maximum $\Delta views$. Figure 9 shows that over 90% videos reach the peak of $\Delta views$ within 7 days. We also observe that after reaching the peak, the $\Delta views$ decays very fast, which confirms the result in Section IV-B1.

V. Conclusion

In this paper, we study the popularity characteristics of UGC associated with hot events mainly based on our 28-day video traces of the largest UGC sites in China. We find that videos about hot events contribute a large number views and should be carefully considered when designing future UGC sites. High skewness in uploaders and video popularity shows great technical and commercial opportunities (e.g., advertisement auctions for top 5% popular videos). Our extensive analysis show that videos associated with hot events only have a very short popular period. The efficiency and information filtering of video search engine might be a limitation of data collection. In the future, we will collect data from Youtube for comparison and establish simple models for predicting popularity evolution, which can provide guidelines of dealing with flash crowd of users during a hot event.

References

[1] http://www.comscore.com/Press_Events/Press_Releases/2010/4/comScore_Releases_March_2010_U.S._Online_Video_Rankings.

[2] http://www.olympic.org/Documents/IOC_Marketing/Broadcasting/Beijing_2008_Global_Broadcast_Overview.pdf.

[3] http://www.cctv.com.

[4] http://www.cnn.com/2009/TECH/06/26/michael.jackson.internet/index.html.

[5] http://www.youku.com/.

[6] http://www.alexa.com/.

[7] http://6.cn.

[8] Youtube: http://www.youtube.com.

[9] http://www.sina.com.

[10] http://www.mop.com.

[11] http://www.useit.com/alertbox/participation_inequality.html.

[12] M. S. Allen, B. Y. Zhao, and R. Wolski. Deploying video-on-demand services on cable networks. In *Proceedings of ICDCS '07*, Washington, DC, USA, 2007.

[13] K.-H. Bae and G. A. Karolyi. Good news, bad news and international spillovers of stock return volatility between japan and the u.s. *Pacific-Basin Finance Journal*, 2(4):405–438, December 1994.

[14] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System. In *ACM IMC*, 2007.

[15] X. Cheng and J. Liu. NetTube: Exploring Social Networks for Peer-to-Peer Short Video Sharing. In *Proceedings of IEEE INFOCOM*, 2009.

[16] R. F. Engle and V. K. Ng. Measuring and testing the impact of news on volatility. Technical Report 3681, National Bureau of Economic Research, Inc, 1991.

[17] C. Huang, J. Li, and K. Ross. Can Internet video-on-demand be profitable. In *Proceedings of ACM SIGCOMM*, Japan, August 2007.

[18] T. Qiu, Z. Ge, S. Lee, J. Wang, Q. Zhao, and J. Xu. Modeling channel popularity dynamics in a large IPTV system. In *Proceedings of SIGMETRICS 09*.