

Understanding the Characteristics of Category-Specific YouTube Videos

Weilong Yang, Zhensong Qian

Abstract—As the world largest video content sharing website, YouTube constantly attracts more and more attention from networking research community. Different aspects of YouTube, such as video characteristics, user behaviors, and its back-end infrastructure have been well studied. Most of these studies consider YouTube as a whole without taking the YouTube category into the consideration. However, it is believed that different categories of YouTube may attract different users and thus have different characteristics. In this paper, we study and compare the YouTube video characteristics from different categories. We found that the videos from different categories have noticeably different statistics on video duration, popularity, user engagement and so on. By directly using the view statistics information provided by YouTube, we have also analyzed the dominant view source from each category, and its growth trend patterns.

Index Terms—Social Media, Networking, YouTube.

I. INTRODUCTION

YouTube, as the world largest video content sharing website, nowadays has become an essential component of the Internet and even people’s daily life. With the emerging of social networks and smart mobile devices, the popularity and impact of YouTube will be boosted significantly. A list of interesting facts of YouTube has been revealed by Google [8], and we highlight a few jawdropping ones here: “48 hours of video are uploaded every minute, resulting in nearly 8 years of content uploaded every day”; “Over 3 billion videos are viewed a day”; “YouTube reached over 700 billion playbacks in 2010”; “YouTube have more HD content than any other online video site”.

When a registered user is uploading a video onto YouTube, he/she will be asked to select one of 15 categories which can best describe the video. In this way, all the YouTube videos can be generally grouped into 15 category. The list of 15 categories and their video distributions are shown in Fig. 1. This figure is from [3] based on the data crawled in 2008. It is shown that both *Music* and *Entertainment* videos are in total 2.5 millions, which is about 50% of all the YouTube videos. Other categories also contribute significant proportion to YouTube. Since videos from different categories will attract different users, we believe they may exhibit different characteristics. Therefore, in this paper, instead of considering YouTube videos as a whole, we measure the characteristics of videos from different categories. Understanding the video characteristics from different categories is very important. From the perspective of networking research, it may help Google or other ISPs optimize YouTube back-end infrastructure or

Rank	Category	Count	Pct.	Pct. (2007)
1	Entertainment	1,304,724	25.4%	25.2%
2	Music	1,274,825	24.8%	22.9%
3	Comedy	449,652	8.7%	12.1%
4	People & Blogs	447,581	8.7%	7.5%
5	Film & Animation	442,109	8.6%	8.3%
6	Sports	390,619	7.6%	9.5%
7	News & Politics	186,753	3.6%	4.4%
8	Autos & Vehicles	169,883	3.3%	2.6%
9	Howto & Style	124,885	2.4%	2.0%
10	Pets & Animals	86,444	1.7%	1.9%
11	Travel & Events	82,068	1.6%	2.2%
12	Education	54,133	1.1%	–
13	Science & Technology	50,925	1.0%	–
14	Unavailable	42,928	0.8%	0.9%
15	Nonprofits & Activism	16,925	0.3%	–
16	Gaming	10,182	0.2%	–
17	Removed	9,131	0.2%	0.5%

Fig. 1. YouTube Video Categories. This figure is from [3]. Note there are 17 categories in the table with two extract ones “unavailable” and “Removed”. We will not discuss these two categories in the paper.

traffic support by providing dedicated designs for videos from different categories.

In this paper, we measure the characteristics of videos from different categories, in terms of video duration, view counts, and user engagement. We found that the videos from different categories have noticeably different characteristics. The dominant view source from each video category is also studied in this paper. An interesting conclusion is drawn that in order to gain the most views in one day, you can either embed your video onto other major (and popular) websites, or get more subscriber for your channels. Besides, we have also modeled the growth trends of YouTube videos. From [3], we learn that the modeling of growth trends of YouTube videos requires the measurements of view counts for a very long period of time (e.g. six months). Besides, it is almost impossible to trace the view counts of a “popular” video from the day one (the time when the video was just uploaded.) Another contribution of this paper is that we use the view count growth data provided in the view statistics HTML subpage of the YouTube video viewing page to analyze the YouTube video growth patterns. We analyze the growth patterns for most-viewed *Music* videos, most-viewed *Music* videos on a day, and recently featured *Music* videos. Those three types of Music videos exhibit very different but intuitive growth patterns.

In the literature, there are many works on characterizing various aspects of YouTube videos. Here we only review the most related works to our paper. Cheng et al. [3] [2] collected two datasets consisting of a few million YouTube videos in early 2007 and 2008 respectively. They provide an in-depth measurement on the YouTube video characteristics, including

video duration, video file size, view counts, video growth trends and so on. More interestingly, they discovered the small-world network phenomena among YouTube videos. In terms of measurement approach, our work is very similar to theirs. Our work can be considered as a supplementary work to their paper. We focus on the measurement and comparison of the characteristics of videos from different categories. In [5], Zhou et al. crawled the view statistics information from the HTML YouTube viewing page. By analyzing this information, they concluded that the dominant view sources for most of YouTube videos are YouTube search and related video list. Our work use a similar measurement on the dominant view source. However, we are more interested in the measuring the dominant view source for each YouTube category.

The rest of the paper is organized as follow. We first introduce our approach of crawling the YouTube video information and our datasets in Section II. Then, based on this information, we analyze the video characteristics in Section III. We analyze the view source of YouTube videos and the growth trends in Section IV and Section V respectively. Lastly, Section VI concludes the paper.

II. YOUTUBE DATA SAMPLING AND COLLECTION

As the world largest video sharing website, it is estimated in May 2011 that YouTube consists of around 400 million videos [1]. There are 48 hours of videos uploading to YouTube every minutes [8]. Therefore, it is almost impossible to crawl all of videos on YouTube. To study the YouTube video characteristics, we have to sample the videos from the YouTube video corpus. However, it is very challenging to ensure the fairness of sampling. For example, the characteristics of newly uploaded videos would be very different to the popular videos. Moreover, the characteristics of videos uploaded from different countries may vary a lot. From a statistics point of view, the more videos are sampled, the generated dataset can better reflect the characteristics of the whole YouTube video corpus. However, crawling a large dataset with a few million videos requires many computing resources and usually takes a very long time. In this section, we will discuss two approaches to obtain the YouTube video information, and how we collected our datasets.

A. Data Collection

To generate our datasets, we use two ways to crawl the YouTube video information, YouTube data API and HTML page. Google has provided a data API which allows a program to search for a video and retrieve its related information [9], e.g. the duration of the video, related video list, category information, user comments, *etc.* Furthermore, YouTube data API also provides the functionality to retrieve a list of videos which reflects YouTube user behavior. For example, YouTube data API can return a list most viewed (or top-rated, most recent, most discussed and so on) videos at today or this month.¹. However, by our experience, the information provided by YouTube data API is still limited. For example, the

view statistics cannot be retrieved by the API. Besides, there is often a constraint on the total number of requests sending to the API server each day. So it is not convenient for us to crawl the individual video information using YouTube data API due to the large amount of videos we need to track and study.

To overcome this limitation of YouTube data API, we also crawl the HTML page of the YouTube video viewing page. More specifically, we use HTML page to collect two sources of information, the view counts, and the view statistics subpage shown in Fig. 2. Note that by our experience, YouTube also poses a constraint on the number of HTTP requests sent from the same IP address, though this constraint might be much more relaxed than the YouTube data API. For example, after sending hundreds of thousands of HTTP requests to YouTube, the requests sent later will periodically lost.

B. Datasets

To study the characteristics of videos from different categories, we retrieve a list of most viewed videos on the day of November 9th, 2011 from each category using YouTube data API. We have crawled 1064 videos in total, and the number of videos crawled from each category is shown in Table II. Due to our limited computing power, we did not crawl more data but rather focus on these 1064 videos in this paper. We believe the measurement approaches proposed in this paper can be easily extended to a large-scale testing video set. Note that for category *Nonprofit* and *Animals*, there are less than 20 videos returned by the API. Although we do not know the exact underlying algorithm to compute the most views video per day in YouTube data API, we believe it is reasonable since both categories *Nonprofit* and *Animals* are less popular than other categories. We understand 1064 videos are negligible compared to the whole YouTube video corpus so that the measurement obtained from this dataset may not well reflect the YouTube video characteristics. However, our goal is comparing the characteristics between different YouTube categories. We believe the measurements from our dataset are comparable across different categories since the videos from each category are sampled on the same day by the same criteria. We have constantly tracked these videos and crawled their related video information from November 9th to December 2th, 2011.

III. VIDEO CHARACTERISTICS

In this section, we summarize the characteristics of category-specific videos, such as video duration, view counts, as well as the user engagement (i.e. ratings, comments, and favorites).

A. Video Duration

The video duration also refers to the length of the video. YouTube is known as a short video sharing website. A 10 minutes video constraint has been imposed on most of YouTube videos. This constraint has been extended to 15 minutes in 2010. Fig. 3 shows the average video durations for each category. Interestingly, as we expected, we can see

¹An example URL we used to request the YouTube data API is http://gdata.youtube.com/feeds/api/standardfeeds/most_viewed_Comedy?v=2&time=today

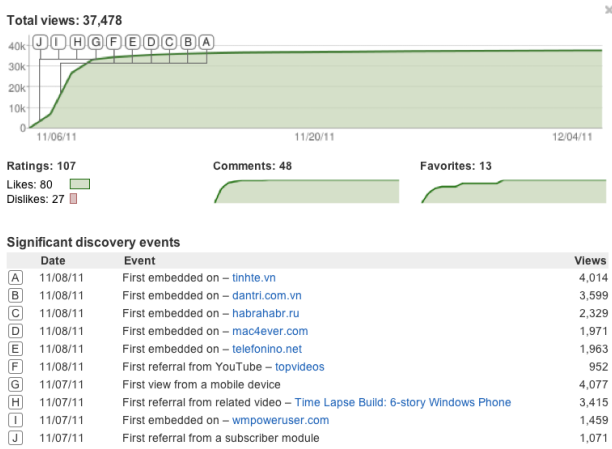


Fig. 2. Snapshot of view statistics. This sub-page can be open by click the “show video statistics” button right next to the number showing the video view counts. The source HTTP page can be retrieved by the URL [http://www.youtube.com/insight_ajax?action_get_statistics_and_data=1&v=\[vid\]](http://www.youtube.com/insight_ajax?action_get_statistics_and_data=1&v=[vid]). (Please change the [vid] to the 11 character videoID you are crawling.)

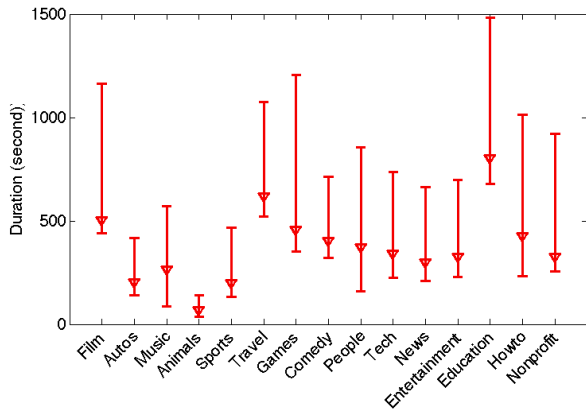


Fig. 3. The average video duration for each category. We have also marked 25%-th and 75%-th duration for each category.

the average video length for music videos is about 4 minutes. Another interesting phenomena is that the average length of animal videos is around 1 minutes. This is very intuitive since most of animal videos are pet videos uploaded by normal users so that they are rather very short. As we can see from Fig. 3, some of *Education* videos are much longer than the 15 minutes limit, that is because some uploader can upload a video longer than 15 minutes through YouTube partnership.

B. View Counts

The view counts is the most direct measurement of video popularity. However, the underling mechanism of YouTube updating the view counts is still unknown to the public. Moreover, in order to avoid spamming view counts, YouTube keeps changing their view count computing algorithms. In our experiments, we learnt that the view counts of YouTube videos are updated every a few hours. Fig. 4 (a) shows the average view counts from each category. As we can see, music and entertainment videos are the top two most popular categories. We have also tracked the average view count gain in ten days for each category, as shown in Fig. 4 (b). Not

TABLE I
THE MAJOR VIEW SOURCE INCLUDED IN THE VIEW STATISTICS PAGE

	source
1	First referral from YouTube search
2	First embedded on <i>some website</i>
3	First referral from a subscriber module
4	First referral from YouTube
5	First referral from <i>Facebook or Twitter</i>
6	Fist referral from related video
7	First view from a mobile device
8	First featured video view

surprisingly, the distribution of average increase of view counts among categories is very similar to the average view count distribution. We can also find out that the Nonprofit is the least popular category which has much less views than others.

C. User Engagement

While a registered user is watching video, he/she can interactive with the video uploader by rating the video (like or dislike), leaving a comment, or favoring this video. This information can be obtained from the view statistics subpage as shown in Fig. 2. YouTube is getting more and more social due to its integration with the social networks. Understanding those user engagement characteristics can help us better model the social network among YouTube users. By counting the average number of ratings, comments and favorites from each category, as shown in Fig. 5, we have observed the following interesting phenomena:

- YouTube users rate more *Music* and *Game* videos than other videos;
- YouTube users leave more comments to the *Howto* videos than other videos;
- YouTube users favorite more *Animal* video than other videos.

IV. SOURCE OF VIDEO VIEWS

Thanks to the embedded links, YouTube video can be found almost everywhere, such as Facebook, blog, and other major websites. We can also easily access YouTube videos through Google search, YouTube video search, or the recommendation sections of the YouTube homepage. It would be very interesting to analysis the source of views for the YouTube videos. As shown in Fig. 2, the video statistics subpage on the YouTube video viewing page shows the source of video views. For example, it shows (entry G in the figure) on the day of Nov. 07, 2011, the given video is first viewed from a mobile device and the views from mobile devices are accumulated to 4077. The major view sources tracked by the video statistic subpage are summarized in Table. I.

By using this view source information provided on YouTube website, we can investigate the dominant view source for a video which contributes the largest proportion of the views of this video. Note that when we compute the dominant view source, we will merge the views from the same type of the sources. For example, assume there are two view sources. One is from “Embedded on website A”, and the other one is

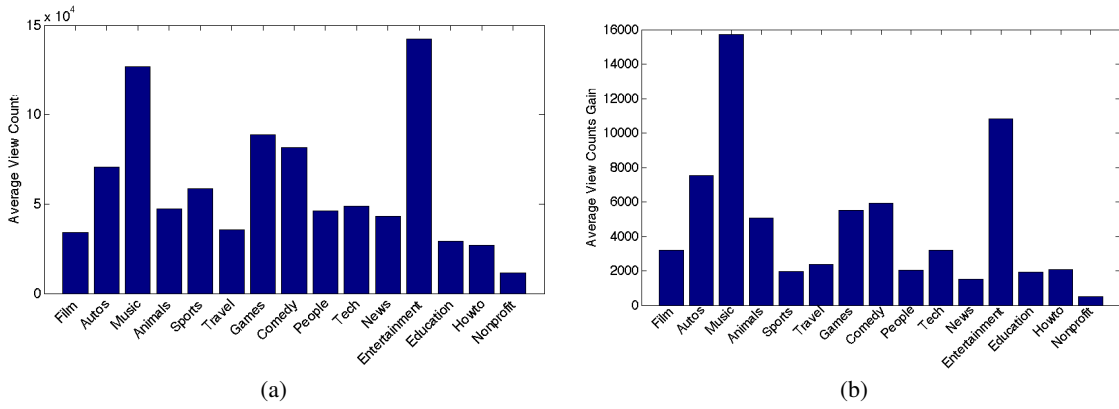


Fig. 4. (a) The average view counts for each category. (b) The average gain on view counts in ten days for each category.

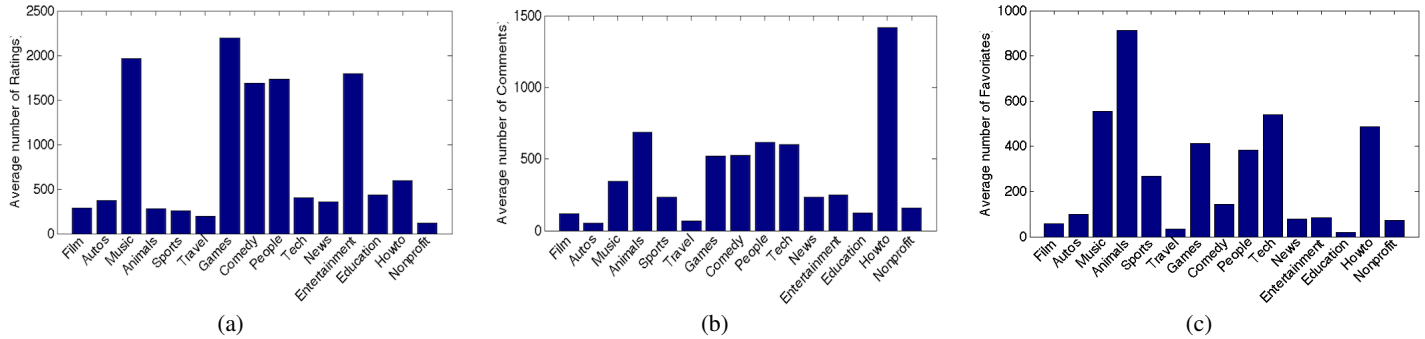


Fig. 5. (a) The average number of ratings from each category (b) the average number of comments from each category (c) the average number of favorites from each category

from “Embedded on website B”. We will consider these two sources as the same source “Embedded” and combine their view counts.

After obtaining the dominant view source for each video, we can easily get the dominant view sources for each category, which are summarized in Table II. We can see the dominant view source of most categories is “Embedded”. For the category *Games* and *Howto*, the most dominant view source is “Subscriber module” which is very intuitive since most of *Games* and *Howto* YouTube channels attract lots of subscribers. Interestingly, the dominant view source of *Music* video is “mobile devices”. This is not surprising because nowadays most of YouTube users like to listen music playlist through their smart phones.

It is worthwhile to note that in [5], their study shows that the dominant view sources for 50% videos in their datasets is related video. However, we believe their findings is not contradict to ours. In this paper, our dataset consists the videos which are most-viewed on a particular day (Nov. 9th 2011). Those videos are usually recently uploaded (less than one month). There is not enough time for them to gain the video views from related video or YouTube search. Our findings can also help answer one question, “How to get the most views on one day?”. Our answer is embedding your video onto other big website, or get more subscriber to your channel if your video is *Games* or *Howto* video.

TABLE II
THE DOMINANT VIEW SOURCE FOR EACH CATEGORY. THE LAST COLUMN SHOWS THE PERCENTAGE OF VIDEOS WHICH HAVE THE GIVEN DOMINANT VIEW SOURCE. FOR EXAMPLE, IT SHOWS THERE ARE 39% OF ALL THE FILM VIDEOS WHOSE DOMINANT VIEW SOURCE IS “EMBEDDED”.

Category	Num. of videos	Dominant view source	% *
Film	94	Embedded	39.4
Autos	43	Embedded	48.6
Musics	74	Mobile Devices	36.6
Animals	19	Embedded	50.0
Sports	98	Embedded	44.7
Travel	15	Embedded	38.5
Games	101	Subscriber module	53.9
Comedy	97	Embedded	29.5
People	93	Embedded	52.2
Tech	87	Embedded	30.2
News	101	Embedded	56.3
Entertainment	97	Embedded	29.5
Education	27	Embedded	40.0
Howto	98	Subscriber module	64.0
Nonprofit	18	Embedded	37.5

V. GROWTH TRENDS

In [3], Cheng et al. has systematically studied the growth trends of the YouTube videos, in terms of the number of views. They updated the number of views of 120 thousand videos every week for 21 weeks. They shows that the growth of view counts of over 80% of YouTube videos will become more and more slowly over the time. This conclusion is very intuitive. But we would rather emphasize that the modeling of

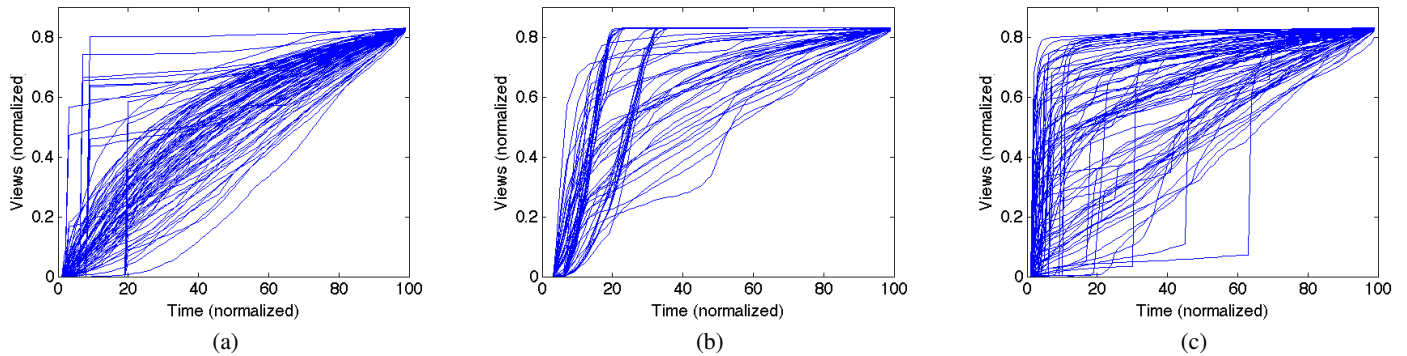


Fig. 6. Growth trends of (a) Most viewed *Music* videos without any time constraints (b) Most viewed *Music* videos on the day of Nov.9th, 2011 (c) Recently featured *Music* videos on the day of Dec.6th, 2011.

the growth trend of YouTube videos is very time-consuming, usually requires a couple of months.

Luckily, as shown in the upper part of Fig. 2, YouTube has already provided the viewing growth trend in the view statistics subpage. Although the growth trend shows as a graph, this graph was generated by Google Chart API [10] so that its corresponding data can be recovered from the HTML source code. We believe this data is very reliable since it is directly provided by YouTube. In [3], the modeling of the growth trend of a video cannot be tracked since this video was uploaded. However, the growth trend information provided by YouTube starts at day 1. This is no doubt a very good news for the growth trend related study on YouTube. We believe many interesting findings can be made by using this data source. To the best of our knowledge, our paper is the first one to analysis YouTube growth trend by directly using the data provided on YouTube view statistics subpage.

After collecting the growth trend data, we can analysis the growth pattern of different types of videos. In this paper, with respect to the *Music* category, we use YouTube data API to crawl three different video lists which reflect different user behaviors, 1) Most viewed *Music* videos without any time constraints. 2) Most viewed *Music* videos on the day of Nov. 9th, 2011, and 3) Recently featured *Music* videos on the day of Dec. 6th, 2011. Each video list consists of around 100 videos. To analysis their growth pattern, we have normalized their growth curve on both time axis and view axis. Note that normalization is done for each video, and we did not normalize the growth data across videos. After the normalization, the growth trend curve can be easily aligned together.

The growth trend curves of these three types of *Music* videos are shown in Fig. 6. These curves are very interesting and intuitive. We have observed the following phenomenas:

- From Fig. 6 (a), we can see the most viewed *Music* videos (without any time constraints) have a roughly linear growth trend. It means those most viewed (i.e. popular) *Music* videos are attracting views constantly. By manually checking the viewing page of those videos, it turns out these videos are indeed the music videos from the most popular singers, such as Justin Bieber and Jennifer Lopez.
- From Fig. 6 (b), the most viewed *Music* videos on the day

of Nov. 9th, 2011, after almost one month, their growth become very slowly, which confirms the observations in [3].

- From Fig. 6 (c), the growth of the recently featured *Music* videos is very sharp. We believe it is because after being featured on the YouTube homepage, those video suddenly attracted a large amount of views.

VI. CONCLUSION

In this paper, we have presented a detailed study on understanding the characteristics of videos from different YouTube categories. The characteristics studied in this paper includes video duration, view counts, user engagement, view source, and growth trends. We have demonstrated that videos from different categories exhibit very different characteristics. These characteristics may introduce more research opportunities on optimize YouTube performance and back-end infrastructure for the videos from different categories. Moreover, we have introduced the analysis of growth trend pattern by directly using the view growth trend data provided by YouTube viewing page. We believe this data is more complete and reliable than crawling and tracking the view counts over a few months. We have analyzed the growth patterns three different type of *Music* videos. These patterns are very different but intuitive.

REFERENCES

- [1] J. Zhou, Y. Li, V. Adhikari, and Z. Zhang. "Counting YouTube Videos via Random Prex Sampling", in *ACM IMC* 2011.
- [2] X. Cheng, C. Dale, J. Liu, "Statistics and Social Network of YouTube Videos", in *IWQoS* 2008.
- [3] X. Cheng, J. Liu, and C. Dale, "Understanding the Characteristics of Internet Short Video Sharing: YouTube as a Case Study", in *IEEE Transactions on Multimedia* 2010.
- [4] M. Cha, H. Kwak, P. Rodriguez, Y. Ahn, and S. Moon "I Tube, You Tube, Everybody Tubes: Analyzing the Worlds Largest User Generated Content Video System", in *ACM IMC* 2007.
- [5] R. Zhou, S. Khemmarat, L. Gao "The Impact of YouTube Recommendation System on Video Views", in *ACM IMC*, 2010.
- [6] H. Yu, D. Zheng, B. Zhao and W. Zheng "Understanding User Behavior in Large-Scale Video-on-Demand Systems", in *EuroSys* 2006.
- [7] X. Cheng, and J. Liu "NetTube: Exploring Social Networks for Peer-to-Peer Short Video Sharing", in *IEEE INFOCOM* 2009.
- [8] http://www.youtube.com/t/press_statistics 2011.
- [9] <http://code.google.com/apis/youtube/overview.html>.
- [10] <http://code.google.com/apis/chart/>