# A Video Sharing Site Oriented Data Collection System

Cheng Wang
Network and Trusted Computing Institute, College of Computer Science, Sichuan University, Chengdu, China
Email: wangcheng8868@126.com

Xingshu Chen and Wenxian Wang
Network and Trusted Computing Institute, College of Computer Science, Sichuan University, Chengdu China
Email: {chenxsh; catean}@scu.edu.cn

***Abstract*—Analysis of video data have significant meaning for studying the characteristics video sharing site and understanding its user's behavior and preference. In order to achieve this goal, we designed and implemented a high performance video sharing site oriented data collection system (VDCS), which overcomes the low efficiency problems that traditional crawler facing. High time efficiency video search strategy is proposed to extract structured video data and high space efficiency filter mechanism is explored for filtering videos. The experiments and analysis proves VDCS have much more practical significance in collecting video data compare with traditional crawler due to its video search strategy and filter mechanism. Through analysis of the datasets we built, we studied video sharing sites' characteristic which includes the upload trend of videos, views and comments and social networks among videos, this study enhanced our understanding of video sharing sites and provide a new measurement tool for analyzing of video data.**

***Index Terms*—video sharing, crawl strategy, filter mechanism, data analysis, social networks**

## I. Introduction

The advent of Web 2.0 and the development of video devices have promoted video sharing, over the course of the past 10 years, many online video websites has emerged in the world, and YouTube [1][2] is world well-known short video network platform. To explore this kind of user-generated content (UGC) system have significant meaning for understanding social public, social trends, hot issues and even users' behavior and attitude on watching video. In such UGC environment, videos are primarily uploaded by users, this pattern makes users are both content consumer and producer. Nowadays, there are many video sharing sites emerged in China, for example YouKu [3] Video and TuDou [4] Video. However, those are not exactly the same pattern as YouTube since its content are not entirely published by users, their videos covering film, TV, animation, variety, documentaries, and other exciting original video. Such difference between UGC system and Chinese Video Sharing Sites inspired us to investigate the characteristics of Video Sharing Sites, obviously that investigate and analysis need large-scale datasets for analyzing. A video sharing sites oriented data collection system here proposed for achieving the goal of collecting large-scale video data from video websites.

For crawling video data from network, many video based search engines have been explored. In [5], the author proposed a video search crawler in distributed environment and discussed key techniques about video search and data collection. Previous works in [6][7][8][9] investigated the method of information extraction, some of researcher also explored extracting structured data from Ajax site, whose data was loaded by triggering JavaScript. All video sharing sites we studied used this dynamical load method. However, those traditional crawlers cannot meet the requirements of scraping videos and videos' related data due to the structured distributed videos on webpage. In [10] [11] [12], authors measured and analyzed the characteristics of UGC system, all of them focused on YouTube. It inspires us to study Chinese video sharing sites, to understand the outstanding features of video sharing sites.

Our works are primarily aim to research and design a high efficiency and comprehensive video-based collection system for obtaining videos and its related information, on basis of this, to analyze and understand the characteristics of video sharing site, saying that how videos grows, transmission and influence people's attitude and behavior. We firstly investigated webpage structure and realized tradition crawler cannot work well in collecting video data, concretely, two reasons are account for it. The first reason is the structural distributions of videos in webpage makes traditional crawler cannot discover and scrape abundant videos in a short running time, specific and high efficiency algorithm need to be proposed. The second factor is that traditional crawler have done relatively poor performance in filtering videos that we have already accessed, however, owing to small world phenomenon [13] exists among videos, it is possible for crawler to process one video not just once. In [14], the author have researched some applications on bloom filter, the experiments showed that the optimal number of hash function minimizes the false prediction.

The highlights of our work are summarized as follows:

doi:10.4304/jsw.8.12.3007-3015

(1) Analyzed video pages for understanding the characteristics of videos' distribution, proposed high time-efficiency video crawl strategy (BFS-DSS). Which largely shorten the time of building our datasets.

(2) We designed high space-efficiency filter algorithm to avoid unnecessary crawl works due to small world phenomena among videos.

(3) Upon completing video data collection system, we conducted our experiment from October, 2011 to May, 2012, during this period we established our datasets. Our analysis works outline the high-level characteristics of TuDou video, which is not only highly representativeness but one of the earliest video sharing site in China, we analyzed TuDou videos' growth trends from TuDou established, besides, video length, categories, views and comments we also studied. Most importantly, we find social networks among video, which have significant meaning for understanding the transmission of online videos.

The rest of this paper is organized as follows. Section 2 outlines system structure, followed by crawl strategy we designed in Section 3. Section 4 studied filtering mechanism and implemented it in our system. Our datasets and analysis results are introduced in Section 5 and in Section 6 we make conclusion of our research.

## II. STRUCTURE OF SYSTEM

Traditional crawlers that focus on text content already cannot satisfy data collection requirements since the specific data that we want to obtain appears structural characteristic, the author of [6][7] also pointed this issue. The needs of designing a high-performance data collection system which takes dynamics data's collection, specific crawling algorithm, and high-space efficiency filter mechanism into account is increasing. Basically, our system contains two parts: Crawler and Extractor. Crawler is used to scraping video links while Exactor plays role in collecting detail information of videos. Figure 1 shows the structure of VDCS. This coordinate work between Crawler and Exactor greatly enhance data collection efficiency.

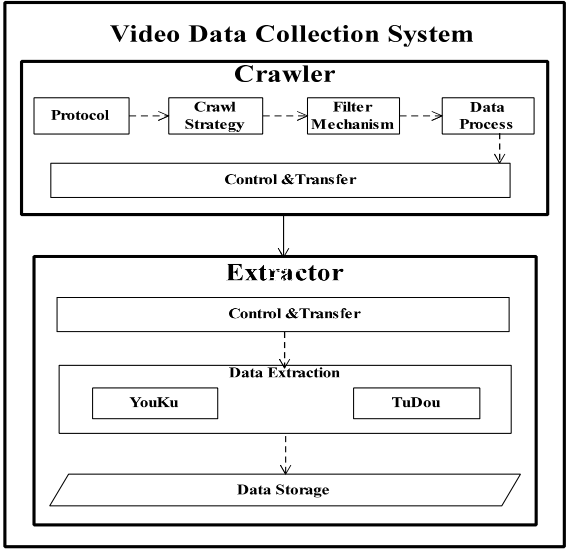


Figure 1.The structure of VDCS

As is shown above, Crawler and Extractor play different role in gathering videos' data. Crawler is mainly focus on scrape video links from each web page, after data processed, the video link data will be packed and sent to Extractor, at the same time, the protocol that Crawler are working for and task types will also be sent to Extractor .Therefore both of them are designed a Control & Transfer module for transferring data .The core module of Crawler is crawl strategy and filter mechanism. In Extractor, data extraction operation focuses on extract some related information of video that including title, tags, description, comment etc.

## III .CRAWL STRATEGY

Before designing crawl strategy, we firstly analyzed the structure of video page, we found almost each video sharing site we studied have several columns pages, interim pages and navigation pages. The analysis shown that there scarcely a video in that kind of pages, but there exist some links that guides us to access webpage which have abundant video resource. According to our analysis, most of videos are exist in secondary or tertiary level. For example, if we access to an video that belongs to one video set (e.g. Series, Album, List etc.), we can find its relative videos, interestingly, it is worth to note that video page limited the number of relative videos to 20, the layered structure of video page appears as Figure 2.

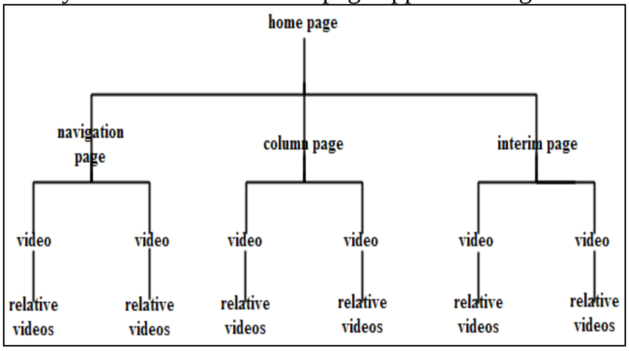


Figure 2. The distribution of videos

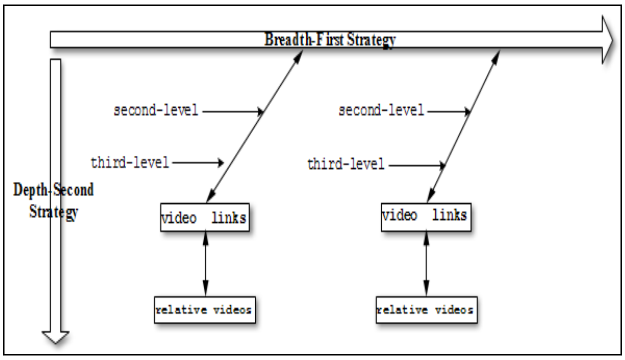


Figure 3. Crawl strategy

Due to layered webpage, some videos aggregated. This distribution makes traditional crawler that used Breadth-First Search [15] spend much more time on discovering and scraping videos. We proposed a video-based crawl strategy, namely, Breadth-First and Depth-Second strategy (BFS-DSS). It means we basically used breadth first search to crawl video links within top-level pages. However if crawler find there some kinds of links that exists plenty of videos on those webpage, the crawler would automatically go to second or third-level page for scraping many relative videos that refer to specific video. After the depth scraping tasks finished, the crawler will return to top-level page continue to do breadth scraping tasks, the whole crawl process can briefly describe as Figure 3, and the algorithm description of BFS-DSS as below.

```
BFS-DSS algorithm description:
Put seed Url to Task queue ;
While(Task_queue is not Null)
 Do
    Get next_request_url←Task_queue.Deueue();
    Access to next_ request_url;
    Extract all video links into link_list
    For i←0; i<link_list.Count;i++
        If link_list[i] have relative videos) Then
            /*Deep down to next level page for scraping relative videos*/
            Insert relative video links into link_list;
        End if
    End for
     Extract all url links into Task queue
  Continue
End while
```

The efficiency of BFS-DSS proved by our experiment, it effectively gained the videos dataset in same run period as BFS, which is shown as Figure 4.In our algorithm performance experiment, we verified the superiority of BFS-DSS algorithm in scraping video from webpage. To simplify description, we named the crawler that implemented BFS algorithm crawler A, and named the other one B crawler, two crawlers are running in same network environment and the same time. In figure 4 our record shown the slope of red line is bigger than black one in first about 3 hours. However, video growth rate of crawler A appears leveling off along with run time‘s increasing. It suggests that there could be more and more pages that contain few videos in task queue from time scale at 100 minutes, in other words, crawler A would crawl those pages in task queue from first to last even through few video in those pages. On the contrary, this phenomenon has little effects on crawler B due to its depth search strategy. It tends to crawl the second or third level pages as long as it finds that there are some related links of individual video in the page. This is mainly reason for constantly growth rate of crawler B.

TABLE I. .
COMPARISON OF TWO STRATEGIES

| Crawl Strategy | Run time (minutes) | Processed pages | Scraped videos | Average videos per page |
|---|---|---|---|---|
| BFS | 245 | 89653 | 282045 | 3.046 |
| BFS-DSS | 245 | 16092 | 334338 | 20.777 |

Another study is about the relationships among run time, number of video and processed pages. Table 1 is the result of strategies in our experiment. For crawler A, the total number of scraped videos and processed pages are 282045 and 89653 respectively, the average number of video per page is about 3.046, In Crawler B that implemented BFS-DSS strategy, it scraped 334338 videos from 16092, average videos per page is 20.777, about 7 times of crawler A. This experiment precisely verified that BFS-DSS strategy is well suitable to scrape videos. As mentioned before, a limitation of displaying relative videos in one page was posed, only 20 relative videos are displayed, thus 21 videos we would collect if our crawler deep down to second or third to collect videos according to BFS-DSS search strategy. The average number of video per page in BFS-DSS is 20.77, which is identical with our webpage analysis.

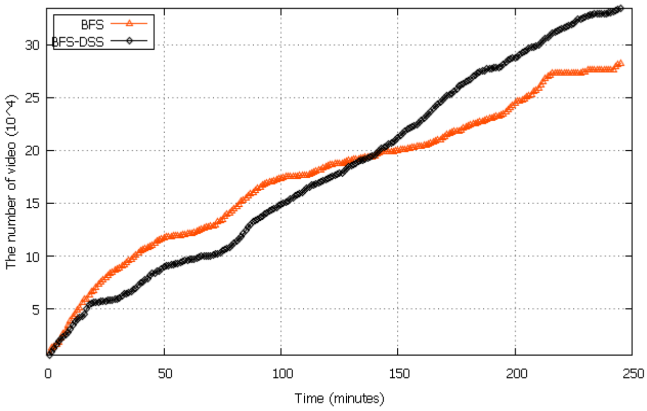


Figure 4. The number of videos along with run time

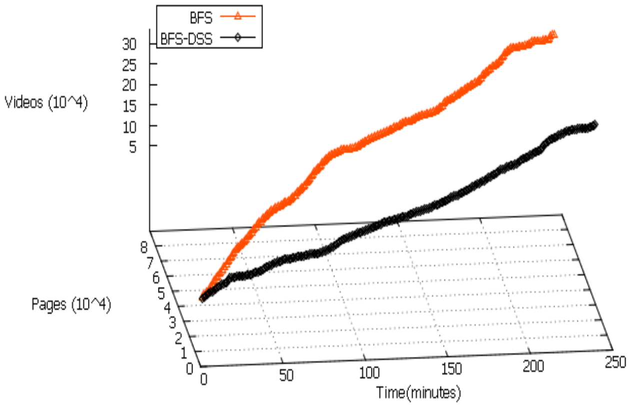


Figure 5. The relation among run time, processed pages and number of videos

Figure 5 is shown as the relations among run time, processed pages and number of videos, in the whole running process, we learnt that the capability of scraping videos in BFS-DSS is better than BFS strategy. Even though BFS strategy gathered more videos than BFS-DSS until 138th minutes, it accessed 57181 pages while BFS-DSS accessed 2770 pages, about 20.64 times of BFS-DSS did. This because more and more interim, column and navigate pages which seldom have videos are added to crawler’s task queue, whereas the crawler A will processed them with no priority. In BFS-DSS crawler, it tends to process those pages have plenty of videos, thus, the capability of finding and scraping videos is better.

## IV. Filter Mechanism

Multitasks program can be paralleled and worked independently. This mechanism makes filter work become a big challenge for crawler, in order to avoid doing repetitive crawling works as less as possible, crawler have to record those pages that have accessed. Our study explored bloom filter found that it perfectly suit to video-based crawler. The Bloom Filter is an algorithm that provides high space-efficient way for checking whether an element is a member of a set [16][17]. As with a hash table, false positives are possible and false negatives are not. Also like a hash table, a bloom filter works by computing hash codes of specific video. The difference is that the bloom filter computes multiple hash codes for each string, and sets multiple bits (one bit per hash code) in the table. The result is counter-intuitive: by setting multiple bits per item, the bloom filter can store many more items in a table of given size while maintaining the low positive rate.

In bloom filter, there have an bit array which contain $m$ elements and all of them are set to 0 at the beginning

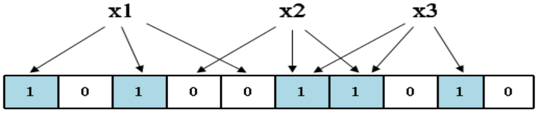


Figure 6. Mapped elements in bit array

Besides, K hash functions are provided to map each element in set $S=\{x_{1,}x_{2,}x_{3,}x_4 \ldots x_n\}$ .Hash function will map every element in $S$ to the bit range $\{1,m\}$ .For example, for each x, if there is exist this element in $S$, hash result will be 1, otherwise, be 0.We can know whether a specific element x is belong to $S$ or not by checking its bit value in bit array, which is mapped by hash functions. Only all bit value that mapped by different $i(0<i<k)$ functions is 1, we conclude this element is in set, or it not belongs to set, in Figure 6, we aim at judging whether element $\{x_{1,}x_{2,}x_3\}$ belong to $S$, we firstly calculate the result of bit value by 3 hash functions. And then check its bit value, we learnt x1, x2 is not belong to set, whereas x3 in this set.

For a given $S=\{x_{1,}x_{2,}x_{3,}x_4 \ldots x_n\}$, which is mapped by $k$ different hash functions to a bit array that has $m$ bits, in this situation ,the possibility of an individual bit that its value still 0 when all element in $s$ are mapped is:

$$p=(1-1/m)^{kn}=(1-1/m)^{-m(-kn/m)}\approx e^{-kn/m} \qquad (1)$$

To simplify our model, we bring an equation: $\lim_{x\to\infty}(1\text{-}1/x)^{-x}=e$ to our model.

In addition, we suppose the ratio of bits that its value is 0 is q, from mathematical expectation formula:

$$E(q)=p \qquad (2)$$

At the same time, M. Mitzenmacher in [17], proved $q$ is intensively distributed near to its mathematical expectation. Thus we can learn the false positive rate $r$:

$$r=(1-q)^k\approx(1-p)^k=(1-e^{-kn/m})^k=k\ln(1-e^{-kn/m}) \quad (3)$$

According to equation (1), (2), (3), we calculate the value of false positive rate:

$r=-m/n\ln p\ln(1-p)$, in order to make sure the minimum value of $r$ is obtained, q has to be 1/2, when q=1/2:

$$r=(2^{-1})^k=(2^{-1})^{\ln2(m/n)}\approx0.6185^{m/n} \qquad (4)$$

We learnt that when q is 1/2, the false positive value is minimum value, from the equation (1), we inferred another equation about:

$$k=-m/n\ln q=-m/n\ln2^{-1}=\ln2(m/n)\approx0.6931m/n \qquad (5)$$

To verify the optimal number of hash functions and what size of bit array we should design for our crawler, we conducted series experiment. From this experiment we learnt the superiority of bloom filter. By given 305565 urls, we tested bloom filter in our crawler from two aspects: memory consumption and positive false rate. In this experiment, we want to make sure the positive false rate of bloom filter is below 0.001,According to our conclusion (4) and (5),we calculated the size of bit array should be at least larger than about 4392496 bits (4.19M)and the optimal number of hash functions should be large than 9.66 . Besides, in order to illustrate the superiority of bloom filter in memory consumption, we allocate same bits size to simple a hash table .As is shown in Figure 7, the collision rate of bloom filter is surprisingly lower than Simple hash table, The 10056 collisions are detected by using simple hash table and 37 collisions for bloom filter.

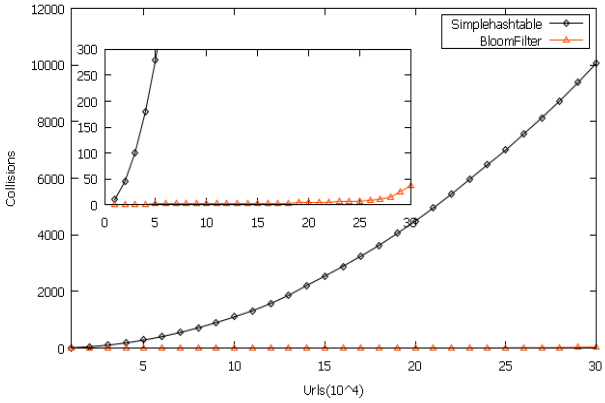


Figure7. Comparison between Bloom Filter and Simple hash table

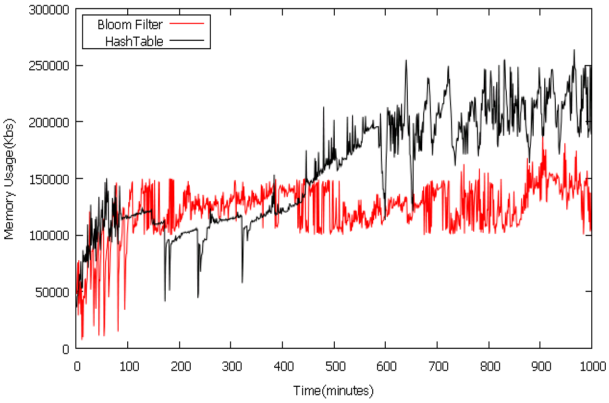


Figure 8. Memory usage of Hash Table and Bloom Filter

Figure 8.is a test about memory usage of two filter mechanisms. The test environment is no other difference except the filter mechanism between crawler A (simple hash table) and crawler B (bloom filter). We ran two crawlers on two different computers, they have all the same configurations and in same network environment. The test period last more than 1000 minutes, in this period we recorded the memory usage of the two crawlers by monitoring program. Obviously, started from 100 minutes the crawler B‘s memory consumption appears relatively steady around 125 Mb. However, crawler A is fluctuated frequently and its memory costs increased along with time, this phenomenon occur due to that more and more urls crawled are added to hash table. Therefore, the occupancy of hash table increased like the Figure 8 shows.

## V. Datasets and Analysis

Based on VDCS we collected massive video data from several aspects, through the analysis of those data we have a depth understand of video sharing sites .In following section we will descript our analysis.

### *A. Categories*

In analyzing video categories of video sharing site, we randomly choose 1.83 million TuDou videos as data sample to study the categories of videos. And we calculated the percentage of each category of total sample videos, the results shows that each video would be classified to one of 16 categories, the distribution of TuDou video categories is shown as Figure 9,according to the pie chart ,Entertainment, TV-play, Music, Animation, Movie, Variety show is the most six popular categories of TuDou video.

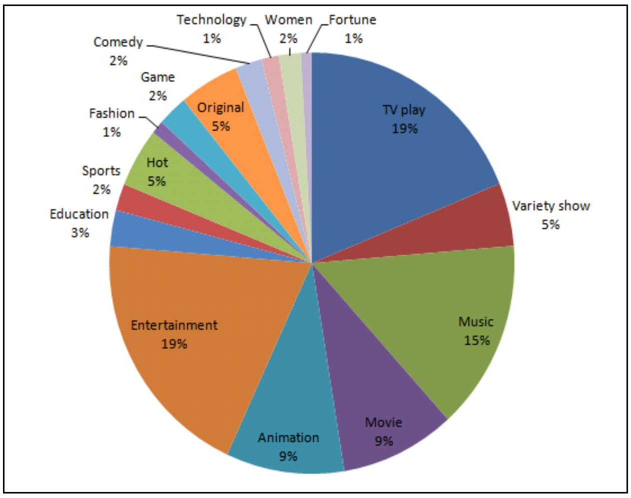


Figure 9.The percentage of each category for TuDou video

### *B. Upload Trend of Videos*

In this part, we examined about 1.83 million TuDou videos we crawled to analyze the upload trend of TuDou videos, during the data collection process, we also recorded the upload date of each video so that we can study the added trend from the time that TuDou system established to May 29, 2012 .Table 2 is our dataset about uploaded TuDou videos, in Figure.10 we plotted the number of new adding videos every month in our sample dataset.

TABLE II.

Uploaded TuDou Videos in Each Year

| Year-Month | Total Number | Min | Max | Median | Average |
|---|---|---|---|---|---|
| 05.4-05.12 | 186 | 11 | 39 | 24 | 20.67 |
| 06.1-06.12 | 21370 | 73 | 5796 | 3046 | 1780.83 |
| 07.1-07.12 | 135626 | 8132 | 13332 | 11570 | 11302.17 |
| 08.1-08.12 | 130470 | 9804 | 14658 | 11948 | 10872.5 |
| 09.1-09.12 | 211864 | 12054 | 24988 | 18034 | 17655.33 |
| 10.1-10.12 | 424646 | 26112 | 51971 | 33816 | 35387.17 |
| 11.1-11.12 | 712080 | 40069 | 103266 | 54140 | 59340.17 |
| 12.1-12.5 | 192517 | 16963 | 90338 | 26541 | 38503.4 |

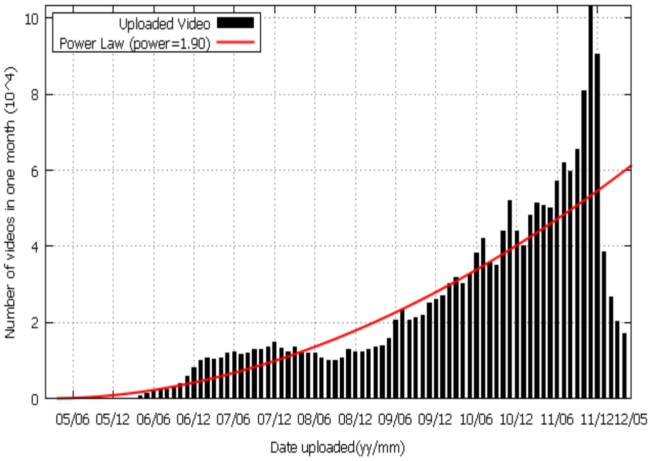


Figure 10. Number of TuDou videos in each month

According to Figure 10, at the beginning of TuDou establishment (TuDou company established in April, 2005), there appears a slow start and a few videos are uploaded by users. However, up to December, 2008, the increase rate appeared relatively steady. Obviously, the uploaded number of videos reached its peak in 2011, it implies that more and more users are willing to upload videos, the trend shown that video sharing sites are become more popular recently years. We also plot a power law curve to fit the growth trend of videos. We find function $f(x) = ax^k$ can well fit our data, we calculated parameter a=12.73 and k=1.90.the red line suggests the growth trend of TuDou video with power 1.9.

On the other hand, we also make statistics of new adding data in each month from April, 2005 to May, 2012 .Table 2 lists the fact of the statistics of uploaded videos in each month. In 2005 ,there are small number of videos are added ,two reason can account for it : (1)part of the videos that uploaded in 2005 have been removed thus our crawler cannot access to them ,(2) most of people are not interested in such video sharing site at that time . The Median and Average number in two periods: 2007 to 2009, 2010-2012 appears relatively steady and increasing constantly. Particularly, the number of uploaded videos in 2011 is highest in Tudou system’s history.

### C. *Video Length*

TuDou video contains both long-content and short content, in this aspect, it is not as same as YouTube video, which limits its video within about 700 seconds .In this part, we used more than 1.2 million videos among we crawled to study the distribution of TuDou video length .Firstly, we divided those videos into 70 datasets, dataset 0 contains the video length between 0 to 100 seconds video. And dataset 70 contains those videos whose video length from 6900 to 7000 seconds, we counted the number of videos in each dataset. Figure 11.we plotted both the facts of video length distribution and CDF to illustrate the video length feature of TuDou video .According to our statistical results. Basically ,there are three peaks in the figure ,respectively ,the first peak is 200-300 seconds ,we assume that most of videos that uploaded by user are relatively short video ,what's more ,according to our categories study ,music video occupied about 15 percent of total videos .As we all known, music video length is usually around 300 seconds, those two factor can account for the first peak. The second peak is 1400-1500 seconds, it possible because most of animation video playtime is about 20-25 minutes .The third peak is 2400-2500 second, which is largely occupied by TV play (19 percent of total video).

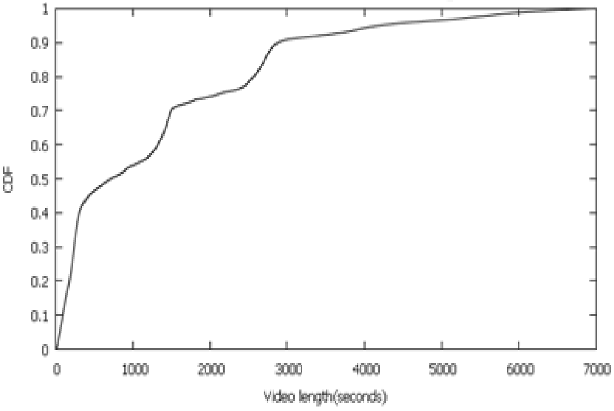


Figure 11.Distribution of TuDou video length

### D. *Views and Comments*

Number of view and comment are significant factors for analyzing video. In this section, we will discuss the some relationships: number of views and rank, number of comment and rank, number of view and number of comment. Through the analysis on these relationships we would have better understanding of user's attitude towards different video types. Our works are also mainly focus on top 100 ranks in top six popular categories like the Figure.12 and 13 shows

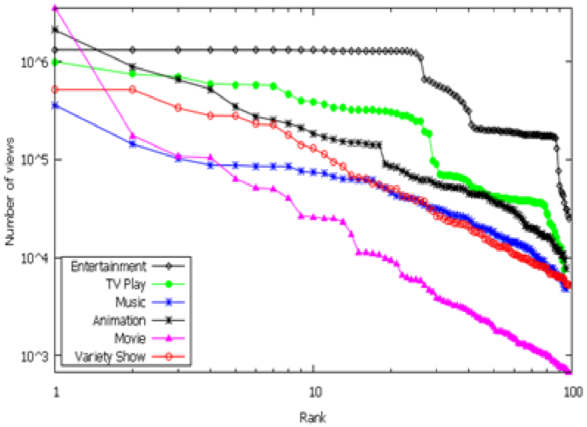


Figure 12.Number of views against Rank of video

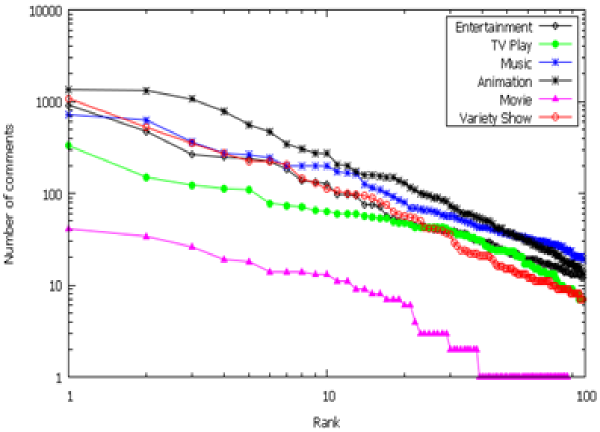


Figure 13. Number of comments against Rank of video

Both views and comments of video is decreasing along with rank, which implies user are prefer to care those hot video no matter news, TV play or entertainment. The number of views of top ten ranks is beyond 1,000,000 in those six categories. Since Movie and TV play is much popular than other categories. Surprisingly, the number of views of most popular TV play reached 1 4,121,908. On the other hand, those video ranking from 20 to 100 is basically posses about 10,000 to 100,000. In addition, Figure 13 shows user more active to express their opinion about an video if it is popular ,on the contrary ,less and less comments are submitted along with rank's increasing. To build the relationship between views and comments of video, we make statistics about the two properties of video, as shown in Figure 14, which suggests most users would not make comments on videos that are not popular, Thus most pots converged below 50 on comments axis, to those playtimes are surprisingly videos, user make more comment and discuss about it.

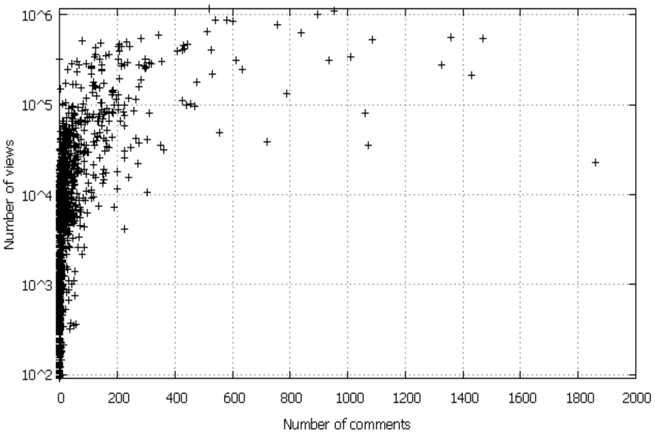
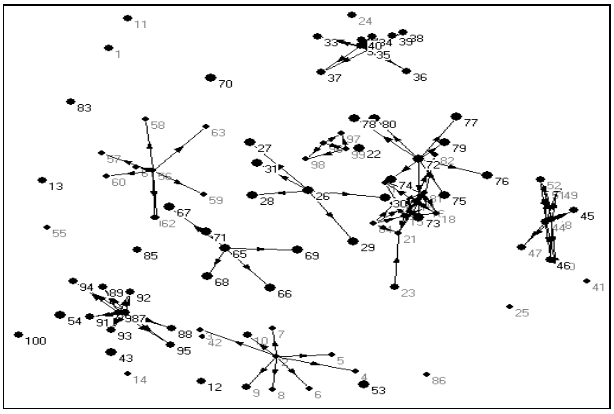


Figure 14.The relationship between views and comments

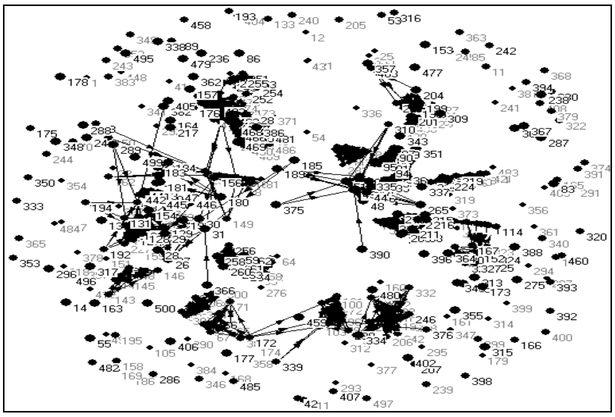
## E. Social Network among Videos

Small world phenomenon is one of most important work for our study video sharing sites. It has been found in several real-world situations [18], for example, road maps, food chains, electric power grids, P2P file sharing network , and Gnutella's research overlay topology. This concept was first introduced by Milgram [19] to refer to the principle that people are linked to all others by short chains of acquaintances, and it was used by Watts and Strogatz [20] to illustrate networks that are neither completely random, nor completely regular, but have characteristics of both. In analyzing small world they defined graph to illustrate the relationship between nodes in social networks.

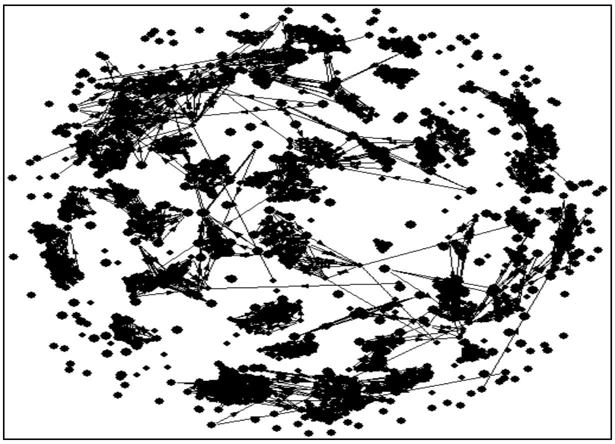
Our study revealed that there exist such clusters phenomenon among videos [21] [22] [23].To study the small world that refer to videos in depth, we recorded the relationship between two videos during crawling process. We defined two relationships between video direct or indirect relation. In video sharing site, if video A and B belong to same set (Album or List) and A can link to B ,we define an edge between A and B.As is shown in Figure 15, some visual illustrations for small world phenomenon among videos, we crawled about 12000 videos and randomly choose about 100,500 and 1000 video scale respectively to study. Because we recorded those direct and indirect relationships between videos, the figure shown that there are some clusters among measured nodes.



100 videos



500 videos



1000 videos

Figure 15.Small world phenomenon among videos

From Figure15 we see that the small world phenomenon of video is obviously by using related videos to form video graph. We also note there are some independent nodes existed, we think this because there are several categories for videos, and some videos have a few or no related videos. Besides, we also learnt other characteristics like clustering coefficient [24][25] diameter of video graph. For a given directed graph G=(V,E) ,which is consisted by several vertices and the line between them ,we define edge $E_{ij}$ is a directed line that connect $V_i$ and $V_j$ ,for a certain vertices k, if the degree of k is deg(k),the clustering coefficient :

$$C_k = \frac{|\{E_{ij}\}|}{\deg(k) \times (\deg(k) - 1)} \qquad (1)$$

$V_i$ and $V_j$ is the neighborhood nodes of $V_k$ .Thus , network average clustering coefficient

$$\overline{C} = \frac{1}{n}\sum_{i=k}^{n} C_k \qquad (2)$$

We used (1) and (2) to calculate network average clustering coefficient for 100,500 and 1000 nodes, we also calculated the diameter of those networks

TABLE III.

SMALL WORLD CHARACTERISTICS OF VIDEO

| Node scale | Average clustering coefficient | Diameter |
|---|---|---|
| 100 | 0.202 | 2 |
| 500 | 0.31 | 3 |
| 1000 | 0.37 | 3 |

We see from table 3, the average clustering coefficient is very high ,especially in scale of 1000 nodes ,along with the increasing of nodes scale ,there is slightly increasing, also, from whole network perspective, we find the diameter of network is increasing slightly, those phenomena suggests that there is a small-world network among videos.

## CONCLUSION

This paper analyzed video sharing sites, find that its highly structured videos data cannot easily collected by traditional crawler efficiently. The distribution of videos inspired us to explore a high time efficiency video search methods, in addition, we suitably designed filtering mechanism for pursuing high space efficiency. The results verified the effectiveness of our works. Such system provides a good platform for our further study in measurement and analysis of video sharing sites. Analysis the video datasets we built enhanced our understanding about the characteristics of video sharing sites, such as growth trend, views and comments that illustrate the nature of behavior in video sharing service, we also explored its social network phenomenon among videos. Those research and analysis have significant meaning for data mining, tracking hot topic and achieving the validity and security of information transmission.

Our future works will focus on video external linking, modeling video transmission, the influences of content distribution network (CDN) and P2P networks to video content distribution and transmission, in-depth study about video's feature in order to suggest video companies provide more enjoyable serv[1]ices for video consumers.

## ACKNOWLEDGMENT

This work was sponsored by the National Key Technology R&D Program (Grant No. 2012BAH18B05).

## REFERENCES

[1] Youtube(Wikipedia)[EB/OL].[2012-12-12] http://en.wikipedia.org/wiki/Youtube.

[2] The Youtube Homepage [EB/OL]. [2012-12-12] http://www.Youtube.com

[3] The Youku Homepage [EB/OL]. [2012-12-12] http://www.tudou.com

[4] The Tudou Homepage [EB/OL]. [2012-12-12] http://www.tudou.com

[5] L.F Yuan. The design and implementation of Distributed Video Search Crawler System[D].Dalian,Dalian University of Science and Technology,2009

[6] Cristian Duda, Gianni Frey, Donald Kossmann,AJAX Crawl: Making AJAX Applications Searchable[C],IEEE 25th International Conference on Data Engineering.2009,pp.78-89.

[7] Xia Tian, Extracting Structured Data from Ajax Site[C],Proceedings of First International Workshop on Database Technology and Applications.2009,pp.259-262.

[8] Kerui Chen, Wanli Zuo, Fan Zhang,Robust and Efficient Annotation based on Ontology Evolution for Deep Web Data, Journal of Computers, vol.6, no.10,2011,pp.2029-2036.

[9] Bo Jiang,Meng-xia Zhu, Jia-le Wang,Ontology-Based Information Extraction of Crop Diseases on Chinese Web Pages, Journal of computers, vol.8, no.1,2013,pp.85-90.

[10] Xu Cheng, K.F Lai, DanWang, and J.C Liu, UGC Video Sharing: Measurement and Analysis [J], Intel. Multimedia Communication: Tech. and Appli., vol.280,2010,pp.367–402.

[11] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez,et al.Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems[J],IEEE/ACM TRANSACTIONS ON NETWORKING,vol.17,n.5, 2009,pp.1357-1370.

[12] Z.Y Li, Rong Gu, G.G Xie.Measuring and Enhancing the Social Connectivity of UGC Video Systems:A Case Study of YouKu[C],Proceedings of 2011 IEEE 19th International Workshop on Quality of Service (IWQoS).2011,pp.1-9.

[13] Adamic, L., Adar, E.: How to search a social network. Social Networks vol.27, n.3, 2005, pp.187–203.

[14] Luo, Qi, Research and Application on Bloom Filter[C], Proceedings of first International Conference on Future Computer and Communication.2009,pp.30-35.

[15] Breadth first search(Wikipedia)[EB/OL].[2012-12-12]http://en.wikipedia.org/wiki/Breadth_first_search

[16] Blooms filter (Wikipedia) [EB/OL]. [2012-12-12] http://en.wikipedia.org/wiki/Bloom_filter

[17] A. Broder and M. Mitzenmacher. Network applications of bloom filters: A survey [J].Internet Mathematics, vol.1, n.5, 2004, pp.485–509.

[18] Social_networking_service(Wikipedia)[EB/OL]. [2012-12-12] http://en.wikipedia.org/wiki/Social_networking_service

[19] Milgram, S.,The Small World Problem. Psychology Today, vol.2, n.1, 1967, pp.60–67.

[20] S. Goel, R. Muhamad, and D. Watts, "Social Search in "Small-World"Experiments", In Proc. ACM WWW'09, Madrid, Spain, April 20 - 24, 2009,pp.701-710.

[21] Wu Peng, Li SiKun. Social Network Analysis Layout Algorithm under Ontology Model. Journal of Software.vol.6, No.7, 2011, pp.1321-1328.

[22] J. Kunegis, A. Lommatzsch, and C. Bauckhage, "The Slashdot Zoo: Mining a Social Network with Negative Edges", In Proc. ACM WWW'09, Madrid, Spain,April 20-24(2009),pp.741-750.

[23] Charu C. Aggarwal, Social Network Data Analytics [M],USA,Springer (2011),pp.19-42.

[24] Xu Cheng,Dale, C.,Jiangchuan Liu, Statistics and Social Network of YouTube Videos[C],Proceedings of 16th International Workshop on Quality of Service.2008,pp.229-238.

[25] Clustering coefficient(Wikipedia)[EB/OL]. [2012-12-12]http://en.wikipedia.org/wiki/Clustering_coefficient

**Cheng Wang,** was born in 1987. He is currently working for Master of Science Degree in College of Computer Science of Sichuan University, Sichuan province, China. He awarded National Scholarship in 2012, University Scholarship in 2011, also awarded Outstanding Postgraduate Student of Sichuan University twice in 2011 and 2012 respectively. His research interests include computer networks, information extraction, network security, web mining and machine learning.

E-mail: wangcheng8868@126.com

**Xingshu Chen,** was born in 1968. She received her M.S. degree from College of Computer Science of Sichuan University, Chengdu, China, in 1999 and Ph.D. degree from Institute of Information Security at Sichuan University, Chengdu, China, in 2004. From 2004 to 2009, she was with College of Computer Science as an assistant professor, working with teaching and research. Currently, she is a professor and Ph.D. supervisor of College of Computer Science. She is currently the director of Network and Trusted Computing Institute. She awarded Scientific and Technological Progress Second-class Award of Sichuan Province of China in September 2008. Her general research interests include peer-to-peer networks, information security, computer networks and cloud computing. E-mail: chenxsh@scu.edu.cn.

**Wenxian Wang**, was born in Jinjiang of Fujian province of China in 1978. He received his M.E. degrees from College of Chemical Engineering, Sichuan University, China, in 2003. He is a lecturer of Network and Trusted Computing Institute, and is currently a Ph.D. candidate at the Institute of Information Security, Sichuan University. He is a member of the International Association of Computer Science and Information Technology (IACSIT).He took part in project “Key Technology of P2P Applications Monitoring System” and awarded Scientific and Technological Progress Second-class Award of Sichuan Province of China in September 2008. His research interests include peer-to-peer networks, information security and trusted computing.
E-mail: catean@scu.edu.cn; catean@163.com.