Optimization of File Allocation for Video Sharing Servers

Emad Mohamed Abd Elrahman Abousabea, Hossam Afifi

Optimization of File Allocation for Video Sharing Servers

Emad Abd-Elrahman and Hossam Afifi
Wireless Networks and Multimedia Services Department, Telecom SudParis (ex. INT), France.
9, rue Charles Fourier, 91011 Evry Cedex, France.
{Emad.Abd_Elrahman, Hossam.Afifi@it-sudparis.eu}

***Abstract*—This paper focuses on one of the most used short video servers on the Internet, it is YouTube. YouTube has the third rank of the internet sites related to its traffic transactions. In this work, firstly, we study the effect of the huge numbers of viewers, hits, users and files on the video sharing servers' behaviours by analyzing some recent statistics about YouTube. Then, we propose an optimization for file allocation procedures in general and then we apply the algorithm to some examples of YouTube videos. This solution improves the number of hits related to those kinds of servers over the Internet. Finally, we try to optimize the revenue from the file allocation and propose a hybrid solution for the file hosting or server caching systems.**

***Index Terms*—Servers hits; Social networks; File allocation optimizations**.

I. INTRODUCTION

NOWADAYS, many servers use short videos as a business model and benefit from a high degree of interest because of their ability to easily diffuse these clips to end users. Users can share their own videos, download available videos and also create their own profiles to know the rating of their videos evaluated by others. Many servers like YouTube [1], Google [6], Yahoo [12], Dailymotion [13] and others have a good ranking as presented by Alexa [8].This evaluation shows the ranking of content depending on uploads and downloads and on the number of videos uploaded by others.

According to the last statistics showed by Alexa, YouTube has a great and increasing interest as it was ranked third site out of the ranked 500 sites with the highest traffic hits over the Internet. We will focus in this paper on this server as a good case study for video sharing servers. Although Yahoo has an advanced rank than YouTube (the second site by Alexa), YouTube is more famous for short videos transfer. Dailymotion also occupies the rank 68 by the same measurement site Alexa.

In this paper, we try to analyse the network architecture of these services and present several enhancements that provide an optimization in important parameters such as response time and storage amount. We show that these optimizations, used at a very large scale, provide really important improvement in the global efficiency of the service as seen by the user and by the provider.

The rest of this paper is organized as follows; section II introduces the history of YouTube and some important statistics, section III discusses the social networking concept and its data centres design, section IV presents our proposal for file allocation optimization, section V shows our algorithm evaluation, section VI draws our attention towards the problem of file hosting on YouTube and its proposed solutions, section VII presents related work and finally the conclusion for our work and its future directions are presented in section VIII.

II. HISTORY OF YOUTUBE

YouTube is four years old, and it is already the third most visited website in the world in 2009. YouTube was founded in February 2005 and it became so immensely popular in a short period of time. It allows people to easily upload and share video clips on www.YouTube.com and across the Internet through websites, mobile devices, blogs, and email. Everyone can watch videos on YouTube, upload, download and also create its own profile on this server to pursue their videos statistics. People can see first-hand accounts of current events and find videos about their hobbies and interests. As more people capture special moments on video, YouTube is empowering them to become the broadcasters of tomorrow [14].

In YouTube's history, USA in 2007 registered the largest number of hits for this site, more than 15,500,000 clicks in the process of measuring the hits of this server related to USA only [2]. In the press of December 2008, YouTube is the leader for online video community that allows people to discover, watch and share originally created videos.

So, now YouTube has more than 100 million viewers, which represents two out of three Internet users who watched online video. Overall, nearly 150 million U.S. Net users watched an average of 96 videos that month. YouTube fans watched some 5.9 billion videos in December 2008[15].By the end of 2008, YouTube accounted for 2.13% of all UK Internet visits compared to 2.11% for Live Mail. During the same week, YouTube was the third most visited website in the UK behind Google UK and Facebook, while Live Mail ranked fourth. And in the recent news for 2009, YouTube.com accounted for 1 out of every 3 U.S. online videos viewed in January. In Germany, 28 million online video viewers watch more than 3 billion videos in December 2008, the statistics in January 2009 indicating that 28.5 million German Internet users viewed a video online in December 2008, up 10 percent versus the previous year [7].

The final press in USA indicates that YouTube Attracts 100 Million U.S. viewers [5].

In brief, the uploaded rate to YouTube was six hours of video every minute in 2007. Then it grew to eight hours per minute, then 10 hours per minute, then 13 hours per minute. In 2008, it became 15 hours of video uploaded every

978-1-4244-6273-5/09/$26.00 ©2009 IEEE

minute. Now, 20 hours of video are uploaded to YouTube every minute, and this is a huge rate and fact the developing of YouTube to become the best online video home. They are dreaming that this site will get 24 hours video uploaded per minute; that means a full day of video uploaded every minute [19].

According to Google's Inc. in Sydney, a report said that; 10 hours of video is uploaded every second to the company's video sharing site [1].

Actually, there is no precise statistics about the number of users or number of hits to YouTube servers available over internet but the majority of this information form the press released on www.youtube.com, (Table III) shows some recent statistics for the grand countries (the most viewed files during this month May 2009) in each country .

We used Google Trends & Analytics [3, 4] to draw some statistics about YouTube servers, hits, uploads and downloads, that will help us imagine the structure of YouTube networks. We also suggest some changes to the structure of this huge social network so as to gain a lot of optimizing bandwidth utilization especially on the international lines between different countries and different continents (as zones).

III. SOCIAL NETWORKING

A. Definition

A social network is a special structure made of nodes representing people or organizations that are tied by one or more specific type of relations such as values, visions, ideas or just friendships. Really, the concept of social networking has been developed more rapidly than the concept of internet itself during the last years.

B. Data Centers Design

In the past, an enterprise might respond to the need for additional capacity and performance with traditional data center infrastructure options (Fig. 1). For example, the enterprises build out the current centralized data center; build additional regional centers when they sign contracts for hosted facilities. Since geography plays an important role in delivering video to global viewers, the design for data centers changes to decentralized one (Fig. 2). Hence, the cost in different world regions must be examined. It will provide a valuable baseline for enterprises like Google or YouTube to compare against alternatives such as Web acceleration services. Therefore, enterprises may consider the expensive option of distributing hardware in regions of the world that are close to users and potential markets. For example, decentralized design is the best solution to build data centers in different zones like; Asia, Africa or with the much heavy loads areas. The files movements in this case may be not expensive if the infrastructure already exists all over the world.

If the enterprises increasingly replace their decentralized data architecture with large centralized data center infrastructure, the capacity becomes an issue. If the data center reaches 95% utilization, the enterprise must find a way to scale its operation to gain more productivity from its current infrastructure [11]. We recommend the decentralized architecture to avoid the server burden problems and obtain a uniform distribution of data centers and not the traditional localized design.

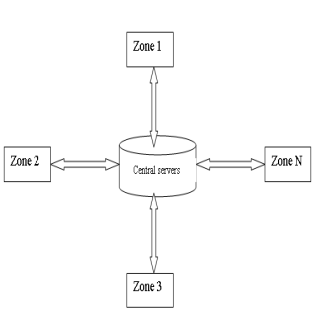

Fig. 1: Centralized Design for Data Center

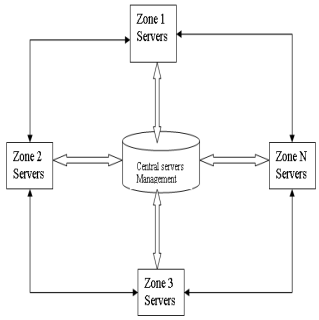

Fig. 2: Decentralized Design for Data Center

IV. THE PROPOSAL FOR FILE ALLOCATION OPTIMIZATION

We present in this section our optimization. We assume that; we have a certain number of zones i (i from 1 to N where N is the max. no. of zones) and a number of servers j (j from 1 to M where M is the max. no. of servers). The zones correspond to the social network consumers and the servers are area servers.

Our goal is to reach a minimum cost and to reduce the number of hits on the main servers. So, we try to have a distribution formula that will lead us to optimize the content locations on the servers in different zones according to minimum cost between zones and servers.

A. Minimum Cost

Let D_{ij} be the cost between zone (i) and server (j) where i varies from 1 to N and j varies from 1 to M. This corresponds to a user from zone i consulting a content from server j. The total cost when consulting the videos in different zones would be $(\mathbf{C_i = \sum D_{ij} * H_{ij}})$ where H_{ij} is the number of hits for a file coming from zone i to server j. we compute $\mathbf{min\ (C_i = \sum D_{ij} * H_{ij})}$ to define the best allocation for any file.

Example, if we have 6 zones and 6 servers (cluster) on principal of one server per zone then:

For zone 1, where the access comes from zone 1:
$\text{Min } C_1 = (D_{11}*H_{11}) + (D_{12}*H_{12}) + (D_{13}*H_{13}) + (D_{14}*H_{14}) + (D_{15}*H_{15}) + (D_{16}*H_{16})$

For zone 2, where the access comes from zone 2:
$\text{Min } C_2 = (D_{21}*H_{21}) + (D_{22}*H_{22}) + (D_{23}*H_{23}) + (D_{24}*H_{24}) + (D_{25}*H_{25}) + (D_{26}*H_{26})$

For zone 3, where the access comes from zone 3:
$\text{Min } C_3 = (D_{31}*H_{31}) + (D_{32}*H_{32}) + (D_{33}*H_{33}) + (D_{34}*H_{34}) + (D_{35}*H_{35}) + (D_{36}*H_{36})$

For zone 4, where the access comes from zone 4:
$\text{Min } C_4 = (D_{41}*H_{41}) + (D_{42}*H_{42}) + (D_{43}*H_{43}) + (D_{44}*H_{44}) + (D_{45}*H_{45}) + (D_{46}*H_{46})$

For zone 5, where the access comes from zone 5:
$\text{Min } C_5 = (D_{51}*H_{51}) + (D_{52}*H_{52}) + (D_{53}*H_{53}) + (D_{54}*H_{54}) + (D_{55}*H_{55}) + (D_{56}*H_{56})$

For zone 6, where the access comes from zone 6:
$\text{Min } C_6 = (D_{61}*H_{61}) + (D_{62}*H_{62}) + (D_{63}*H_{63}) + (D_{64}*H_{64}) + (D_{65}*H_{65}) + (D_{66}*H_{66})$

So, the minimum value of C_1 to C_6 will give us an approximate location for the best allocation for this file demands so as to save bandwidth and time for viewing or downloading. The evaluation results will lead us to logically move this file to the region from which the highest demands are coming. We can follow the steps of the proposed algorithm in Matlab in (Table I).

B. The Correlation between Demographic Information and Web Usage

In this part, we try to find a relation between the network topology and the cost calculated between zones for any file. We have yet calculated the cost related to zone (1) by C_1 and make the same for zone 2 to 6. Then, we can decide the best allocation for this file according to these values but based on the relation between zones:

1). First Assumption: Full Mesh Topology

If $C_1 <$ (C2 to C_6) then the best location of this file is the servers in zone (1), this means that the minimum value C_i will define the best location of that file. Actually, the full mesh design suffers from the N^2 problem where, the number of links needed for this design equal N (N-1)/2, and no one can say that there is a warranty for full mesh design on the internet (see Fig. 3 left-half).

2). Second Assumption: Not-full Mesh Topology

If the network is partially meshed (see Fig. 3 right-half) then, the previous calculations and decisions will be different, and we must take into account the cost of the intermediate zones. For example, if zone (1) is not directly connected to zone (2), but connected through zone (3), then, while making our calculations, we must add the cost between zone 1&3 and 2&3 for the calculation of zone (1).

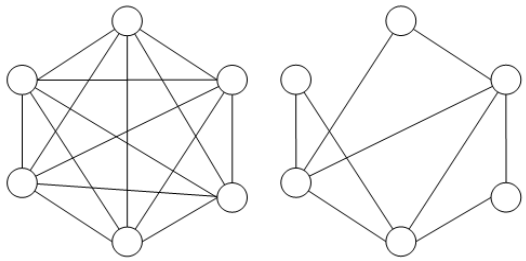

Fig. 3: Full Mesh design for 6 zones & Partially Meshed design

TABLE I
THE PROPOSED ALGORITHM.

Algorithm

Input N (number of zones)

Input M (number of servers)

Input h_{ij} (number of hits come from zone i to sever j for specific file)

Input d_{ij} (the assumed cost between zone i and server j)

Total cost $C_i = \sum hij * dij$ // where i from 1 to N and j from 1 to M

If C_i min than C_{i+1} to C_N

Then allocate this file in zone (i)

Else moves this file to the Min (C_{i+1} to C_N) value location

End If

Return the best location for this file (best i)
End

V. THE PROPOSAL EVALUATION

In this section, we focus on the evaluation of our algorithm. We choose some files randomly from YouTube site that rated as top viewed during this period. We select 6 files from different zones as listed in table II, where each file has its own statistics related to hits and users from different zones. These statistics are estimated by the percentage of YouTube users for each file according to the file profile on YouTube site and the statistics from Alexa [8] for YouTube users/zone utilization.

We apply the algorithm in table I on the six files selected in table II file by file and we take into consideration the different locations of the file. Therefore, we run the algorithm 6 times for each file by assuming the movements of the file from zone to another and optimize the best location according to the minimum cost equation (**min ($C_i = \sum D_{ij} * H_{ij}$)**).In Fig. 4, we illustrate the results for the files distribution. We notice that the hits of the six files in the different zones play the big role in the minimum cost calculations.

For file 1, its best location is the servers in zone 1 where the minimum cost of that file investigated by the algorithm. **For file 2**, it must be moved from its location in zone 2 to zone 3 where the algorithm gave the minimum cost. For the rest of files shown in Fig. 4; **file 3** moves from zone 3 to zone 2, **file 4** moves from zone 4 to zone 3, **file 5** remains in the same zone 5 for achieving minimum cost and finally **file 6** moves to zone 4.

We conclude that, we can apply that algorithm on any file for which we have some statistics to define its best location on the video sharing servers.

TABLE II
EXAMPLES OF SOME FILES CHOSEN FROM YOUTUBE AND RELATED HITS

Files/ Hits	File 1 Hits 425783	File 2 Hits 51340	File 3 Hits 174917	File 4 Hits 27797	File 5 Hits 12403	File 6 Hits 136394
Zones/ %users Access YouTube	Zone 1 users 30.95 %	Zone 2 users 9.25 %	Zone 3 users 30.85 %	Zone 4 users 23.05 %	Zone 5 users 2.45 %	Zone 6 users 3.05 %

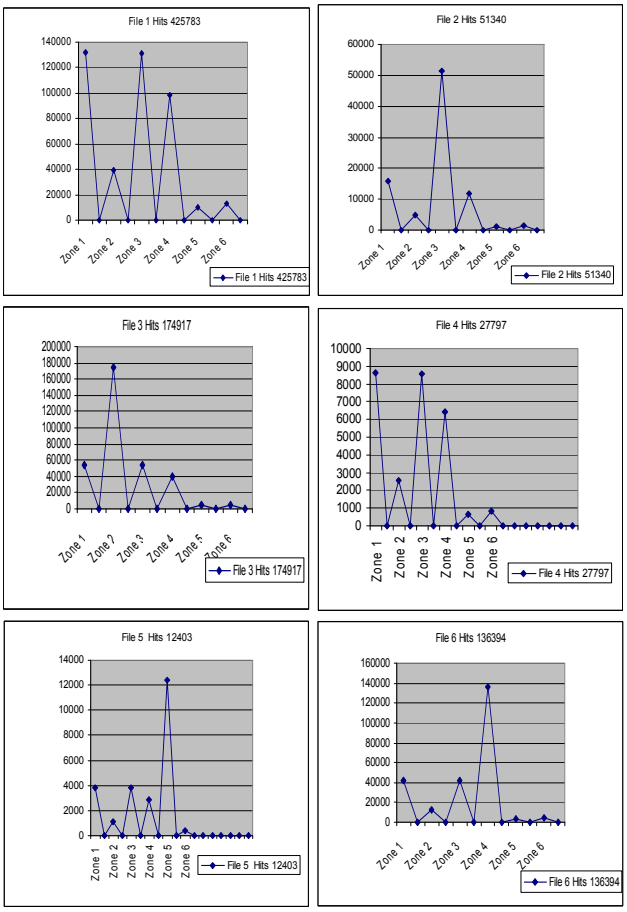

Fig. 4: Files distribution by their hits and the cost related to zones.

VI. THE FILE HOSTING PROBLEMS ON YOUTUBE

We claim that YouTube popularity will cause a huge storage problem because the number of files will increase without any revenue from this hosting**.**

A recent report [17] about YouTube video growth curve assures that YouTube currently streams more than 30 billion videos per month worldwide. This huge number of videos drew our attention towards studying the problems of inflation of the server database and how the company can benefit from the hosting of those files. The challenge is how YouTube will treat this difficulty without affecting the online video traffic marketing (there is a forecasting that; by the end of the year 2009, YouTube will lose about 470 million Dollars and the company will need to change its strategies [18]).

A lot of analyses on the continuing growth of YouTube's library versus the cost of storing files were made. Since the storage of video files that are not viewed will grow and should it be YouTube responsibility to keep these video records for the future and if so can the sponsors support additional cost for the long time hosting of those files? We believe that, Google needs to study the phenomena of this growth and think how they will find a good solution for managing this problem. YouTube could generate revenues by changing its strategy towards gaining money that will not rely only on advertisements but also from hosting files. We think that Google will gain from the investment that will return from files hosting on YouTube servers.

A. The Proposed Solutions for Video Hosting Costs

In this part we have two problems to solve; the first one is the long term hosting of files that become old and the hosting cost in relation to the revenue from hosting:

1). Long Term Hosting and the Hybrid Algorithm

We can add two flags numbers related to each added file on the server;

- The first one is the history flag (F_h) that can register the history of this file by counting the number of days this file allocated or uploaded to the server.
- The other flag is (F_v) which represent the number of viewers of that file.

By calculating the ratio (F_v/F_h) in a certain period of time, we will have an indication that this file deserves to remain on the server in case the ratio is high or to be omitted from the server database in the other case. This ratio expresses the File Time-To-Live FTTL on the server, which is given by FTTL = (F_v/F_h). Most Webs caching systems depend mainly on the time period which the files spent on the servers for the decision of continue hosting or deleting (Least Recently Used (LRU) idea). Others depend on the number of views for those files (Frequency Based) approaches. But in our proposal we used the two parameters (times & numbers) for calculating FTTL. **So, we can consider this algorithm as a hybrid algorithm for manage caching or storing files.**

2).The Relation between File Hosting Cost and its Revenue

In this paragraph, we study the behaviour of file hits on their hosting servers and the revenue returned from hosting this file or moving to another zone server according to the file allocation algorithm in table I.

The following parameters will be used in optimization of the total revenue from hosting files;

H_{ij}: Numbers of hits/day for a file on server j, accessed from users in zone i.

D_{ij}: Cost of hitting a file on server j, from user in zone i.

$C_j = \sum_i H_{ij}.D_{ij}$: Total cost/day for a file put on server j.

$H_j = \sum_i H_{ij}$: Number of hits/day for a file on server j.

$\bar{C}_j = \frac{C_j}{H_j}$: Average cost/day for a file put on server j.

$\bar{R}_j$: Average revenue generated by a file put on server j.

$R_j = H_j.\bar{R}_j$: Total revenue generated /day for a file put on server j.

The formula to compute the Revenue Cost Ratio (*RCR);*

$$RCR = \frac{R_j}{C_j} = \frac{\bar{R}_j.H_j}{\sum_i H_{ij}.D_{ij}}$$

This ratio will give us a good indication for the revenue from hosting this file in terms of its hosting cost with relation to the file hits.

Finally we can calculate the net revenue by the following equation;

$$R_j - C_j = H_j.\bar{R}_j - \sum_i H_{ij}.D_{ij}$$

And this revenue is considered as a good indication also for the decision of hosting the file or not.

TABLE III
STATISTICS & EVALUATIONS FOR THE MOST FAMOUS COUNTRIES BY YOUTUBE.

COUNTRY	HITS/DAY FOR THE MOST VIEWED FILE	% OF USERS ACCESS YOUTUBE(X)	INTERNET USERS (MILION)(N)	YOUTUBE VIEWERS/ MONTH	TOTAL TIME(MIN) (AV = 22.9 M/D/U)	PROBABILITY OF YOUTUBE USERS(X/N)
Australia	136394	0.6	17	102000	2335800	0.035
UK	174917	3.8	43.8	1664400	38114760	0.087
Canada	5710	2.3	28	644000	14747600	0.082
Ireland	117914	0.6	2.4	14400	329760	0.250
Brazil	26807	3.8	67.5	2565000	58738500	0.056
France	51340	3.7	40.9	1513300	34654570	0.090
Japan	27797	7.7	94	7238000	165750200	0.082
India	28667	4.0	81	3240000	74196000	0.049
USA	425783	22.8	220.1	50182800	1149186120	0.104
Germany	29414	4.7	52.2	2453400	56182860	0.090
Spain	40291	2.6	28.6	743600	17028440	0.091
Russia	20601	1.4	30	420000	9618000	0.047

B. Traffic and User Social Behaviour

From table III, we can see that; the traffic of YouTube has a great effect in attracting the viewers. We noticed in one hand, the average time for each user spent per day is 22.9 minutes. This time is not a small period if we make comparisons with other sites like news or mails sites. Therefore, all internet researchers expect this time will increase dramatically in the near five years. On the other hand, a lot of countries which have a large percentage of users accessing YouTube like USA not by default give us an indication that it is the country that has the maximum percent of utilization in terms of the probability of YouTube users. We need to take into account the total number of internet users as a base for that calculation as the results appeared in the last column of the table III above.

The social relation between YouTube viewers and videos is a significant relation but, we can not put it under any mathematical equation. The number of visitors to YouTube as it appeared in the table III has no relation if we compared it with the last column of the table (the probability of YouTube users(x/n)). For example, if we compare USA with Ireland we can find that; the viewers' probability for Ireland is 25%:10.4% to USA. Although the number of internet users in USA is 220.1 M: 2.4 M to Ireland which almost ratio 100:1.

VII. Related Work

Many solutions have been proposed for file allocation for videos and other heavy loaded servers and data over the Internet by cashing files like AKAMAI systems [11], which facilities a lot for video delivery and streaming cashing solution. But, the research still tries to find solution design for servers and files allocation procedure with low cost.

Many measurements and statistics, like the study in [9] have been carried out to show the grand utilisation of social networks and the increasing users numbers uploaded or downloaded videos through internet using famous sites like YouTube or Dailymotion. Those latter are the most famous sits. When you come to understand their traffic pattern, YouTube is a significant web site even among all web 2.0 sites. This is due to the fact that content of YouTube is video that consumes much more bandwidth than text, picture and audio. Some statistics regarding the network of the University of Calgary and YouTube would be useful. The university has 28000 students and 5300 faculty and staff. The data for this study was collected in 85 consecutive days in spring 2007. It turns out that YouTube traffic constitutes 4.6% of the whole internet traffic of the university. YouTube is the most popular video-sharing web site and it is the source of 60% of the videos watched over the internet. 10,000,000 videos are downloaded to watch every day and 65,000 new ones are uploaded.

There is another work which presented a systematic and in-depth measurement study on the statistics of YouTube videos [10]. When authors analyzed these statistics, they found that YouTube videos have noticeably different statistics from traditional streaming videos, in aspects from video length to access pattern. They also studied some new features that have not been examined by previous measurement studies: the growth trend and active life span of videos.

In [16], the authors tried to analyze the video and the user characteristics for different geographical regions, concentrating mainly on Latin America. They developed an efficient way for collecting data about videos and users. Based on the collected data, they showed that there exists a relationship between geography and the social network features available in YouTube. They presented evidence that indicates that geography creates a locality space in YouTube, which could be used to explore infrastructure improvements, such as caching mechanism and content distribution networks.

VIII. Conclusion

The field of social network design and short videos delivery like YouTube becomes an important research domain.

In this paper, we tried to show the history and analyze some statistics related to one of the most popular video delivery; YouTube.com. We also focused on the social network design for such types of servers and presented an optimization for file allocation that gave us two contributions. The first contribution moves files from server to server according to the minimum cost between zones and servers. The second contribution links hits on some servers with file allocation and geographical distribution of servers. We also optimized the revenues from file hosting and file movements.

As a future work, we hope to optimize the allocation of files and their distribution in case of Peer to Peer networks under complex conditions like the hidden or out of work of one node in the logical network for Peer to Peer design.

IX. References

[1] YouTube Members: http://www.youtube.com/members.
[2] YouTube Press: http://www.youtube.com/press_room.
[3] Google Trends: http://www.Google.com/trends.
[4] Google Analytics: http://www.Google.com/Analytics.
[5] Http://www.webpronews.com/topnews/2009/03/05/youtube-attracts-100-million-us-viewers.
[6] Google Blog: http://www.google.com/blog.
[7] ComScore Press: http://www.comscore.com/press.
[8] Alexa: http://www.alexa.com/.
[9] P.Gill, M.Arlitt, Z.Li, A.Mahanti, 'YouTube Traffic Characterization: A View from the Edge', ACM Internet Measurement Conference (IMC) San Diego, CA, USA October 2007.
[10] X. Cheng, C. Dale, J. Liu, 'Statistics and Social Network of YouTube'. IWQoS 2008. 16th International Workshop on Videos, Quality of Service, IEEE 2008.
[11] AKAMAI : http://www.akamai.com
[12] Yahoo: http://www.yahoo.com
[13] Daily Motion: http://www.dailymotion.com
[14] Http://mediatedcultures.net/ksudigg/?p=108.
[15] PC World: http://www.pcworld.com/article/158949/
[16] F.Duarte, F.Benevenuto, V.Almeida, J.Almeida, 'Geographical Characterization of YouTube: a Latin American View', Web Conference, 2007. LA-WEB 2007. Latin American, Oct. 31 2007-Nov.2 2007 Page(s):13 – 21.
[17] 'YouTube currently streams more than 30 billion videos per month worldwide',http://www.fierceonlinevideo.com/
[18] Http://www.multichannel.com/article/191223YouTube_May_Lose_470_Million_In_2009_Analysts.php
[19] '20 Hours of Video Uploaded Every Minute', http://www.youtube.com/blog.