A STATISTICAL AND CLUSTERING STUDY ON YOUTUBE 2D AND 3D VIDEO RECOMMENDATION GRAPH

Ioannis Tsingalis, Ioannis Pipilis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Greece
{itsingal,pitas}@aiia.csd.auth.gr

ABSTRACT

Social network sites, like Facebook, Tweeter and Flickr provide users the opportunity to share their media content, such as videos, music tracks or photos. Beyond the fact that they can share information the users can also vote or comment on information posted by other users. Social networks take advantage of this activity and create groups and communities of users with similar interests. This categorization helps social network systems to support users with data, e.g. videos, photos or users profiles, that are relevant to their interests. In order to increase the effectiveness of navigation, analysis of social media content graphs needs to be done. In this paper, an analysis of the Youtube social media graph is presented. Graphs of 2D and 3D videos are considered in this analysis. Well known properties of web and social networks analysis, like the power-law distribution are discussed. Moreover, clustering methods are applied in order to study the existence of media content groups. Finally, the results of our analysis are discussed and directions of future work are presented.

Index Terms— YouTube recommendation graph, 3D, social networks, analysis, clustering

1. INTRODUCTION

Graph social network analysis is fundamental for the improvement of the functionalities of the already existed social networks, but also for the establishment of new ones. Interesting studies about the structural properties of social networks, such as the local clustering, small-world behaviour, power-law distribution and scale-free properties have been presented in the past [1–4]. More specifically, Mislove et al. [1] worked on large scale datasets that they have collected from the most popular social network sites, e.g. YouTube, Flickr, Oktur and LiveJournal. Using structural properties, like small-worldness, power-law distribution or scale-free behaviour, it was found that each social network exhibits a different structure in the web. More precisely, it was discovered that social networks are characterized by a large number of small tightly clustered communities. Similar observations about small-world behaviour and local clustering have been presented by Adamic et al. in [2]. They also studied distributions of video length, video categories and graph structural properties such as small-worldness. In another research on the social networks [3] it was shown that the average path length between two Americans is 6 hops. Ahn et al. [4] studied three social network services, namely Cyworld, MySpace, and Orkut. They used metrics like the degree distribution and clustering coefficient and they found that Cyworld has a multi-scale behaviour while Myspace and Orkut follow the power-law distribution. They also studied the evolution of the social network structures over time by collecting datasets in different periods of time. For a comprehensive analysis of social networks one can refer to the book by Wasserman and Fraust [5].

In this paper, we concentrate on the YouTube content graph analysis. YouTube was established in 2005, and since, then it has become the 3rd most accessed site on the internet, after Google and Facebook. In 2012, YouTube has stated that four billion videos were served per day. An interesting feature, which is offered since 2009, is that users can upload two channel stereo videos, called 3D videos in contrast to the classical single-channel(2D) videos. Thus, 3D viewing experience can be provided to the users. YouTube flash player can display anaglyph videos in red/cyan, green/magenta or blue/yellow layout. Also row/column interlaced display is also provided when a YouTube video is displayed on the screen. YouTube also presents a list of 'relevant' videos. Usually, the relevance is directly related to the order in which a videos appears in the list, the first ones being the most relevant to the video on display. In this way, a recommendation or relevance graph of YouTube videos can be created, where videos are graph nodes and graph edges denote relevance of two videos. Initially, the graph edge weights are assumed to be equal to one. In the rest of the paper we will call the *recommendation* or *relevance* graph just as graph or video graph.

The rest of the paper is organized as follows. The statistical study of YouTube recommendation graph is demon-

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

strated in Section 2. In Section 3 clustering experiments are presented. Conclusions and future work are summarized in Section 4.

2. YOUTUBE STATISTICS

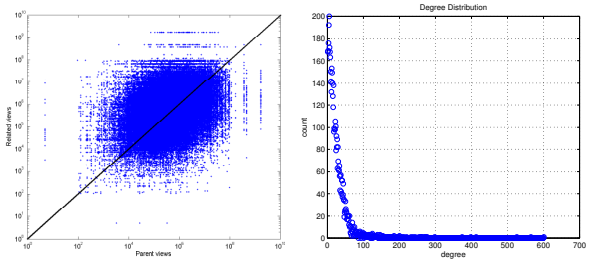
Crawling social networks and, more specifically, YouTube is a challenging task because of the amount of information that has to be traversed. That is why we restricted the analysis to a sub-graph of Youtube. Breadth First Search (BFS) and Depth First Search (DFS) are the most common algorithms for web crawling [1]. Both of them have their advantages and disadvantages. Our Crawler is based on the BFS method, also known as snowball method. In the snowball method, we start the crawling from one node and collect the nodes that are connected with this first node making a second layer of nodes. The same method is applied to the nodes of the second layer. This procedure is repeated, until we collect a specific number of nodes. With this method the data seem to increase like a snowball. Starting the snowball method from different nodes does not affect the clustering coefficient. Unfortunately, the same task can affect the power law distribution as we see bellow. For more information on this subject one can refer to [6] and [7].

We structure the information we collect via the YouTube API into the recommendation graph described in Section 1. Each graph node refers to a video and contains relevant information, e.g. video title, views, number of likes and dislikes. An edge in the graph refers to the connection between a parent video and the related videos that API returns for a parent video. A weight to each edge is assigned according to the order the related videos are returned. YouTube API was configured to return the fifty relevant videos. This is the maximum number of related videos that the YouTube API can return.

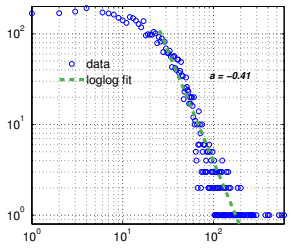
We collect two graphs. The first has 5000 nodes with 3D and 2D videos. In order to ensure that the first graph contains both 3D and 2D videos we start the crawling from a 3D video. The second recommendation graph has 4000 nodes with only 3D videos. In order to obtain the 3D recommendation graph, we also start crawling from a 3D video and when a 2D video is returned it is not included in the graph. We call the graph containing 3D and 2D videos *unfiltered* graph, while the graph that contains only 3D videos is called *filtered* or 3D video graph.

Using the views of each video in the unfiltered graph, we can plot the relation of the number of views between a parent video and the number of views of a relevant video (Figure 1a). In there the X-axis represents the number of views of a parent video, while the views of the relevant videos to the parent video are depicted on the Y-axis. We observe in Figure 1a that most of the points are concentrated along the diagonal. This means that, for a YouTube video with a specific number of views has related videos with similar number of views. In other words, videos are proven to be grouped according to their popularity. An other interesting observation is that, the video pairs are not symmetric to the diagonal line, since there are more points above the diagonal. This means that, it is more likely that YouTube will recommend a more popular video than the one we are already watching. The opposite is less likely to happen.

The degree of a graph node is defined as the number of links that are attached to it in the graph. Using the degree of each node, we can study the structure of the graph, by considering the node degree distribution (Figure 1b). We observe that the degree distribution has long tail characteristics. This means that most of the nodes in our graph have low node degree, which in turn means that YouTube mostly consists of videos with medium popularity.



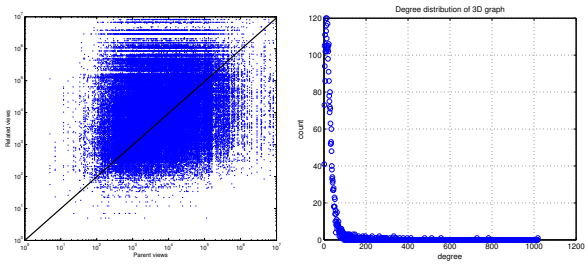
(a) Views of related videos (b) Degree distribution



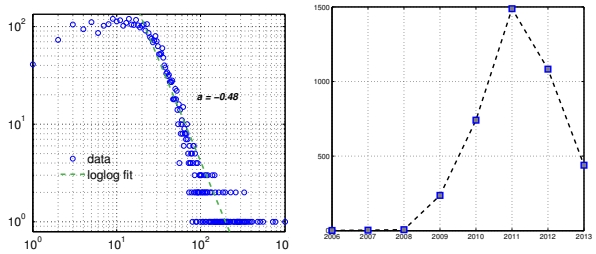
(c) Log-log scale distribution

Fig. 1. Statistics for the unfiltered graph

Usually, in social media and web analysis we study if the degree distribution follows a power law one [1–4]. In other words, we study if a network has free scale characteristics. The log-log transform of the degree distribution is pictured in Figure 1c. Namely, in power-law networks the probability a node to have degree k is proportional to k^a. The power law in log-log coordinates graph a straight line. The line with slope a is computed for both unfiltered and 3D graph (Figure 1c and 2c), without taking into consideration nodes with degree smaller than 25. The flat head of the distribution with degree smaller than 25 is due to fact that the BFS under-samples low degree nodes [6]. More specifically, the last layer of the BFS tree which contains mostly low degree nodes, this is verified from the long-tail degree distribution, is not complete because the crawler has to stop when the desired number of nodes are collected.



(a) Views of related videos (b) Degree distribution



(c) Log-log scale distribution (d) Uploading rate

Fig. 2. Statistics for the 3D graph

The degree distribution of the graph that contains only 3D videos and the corresponding log-log transform of the degree distribution are depicted in Figure 2. We observe similar characteristics to the ones we discussed for Figure 1, which refer to the graph that contains both one and two channel videos. Finally, for the 3D graph we plot the uploading videos distribution for the period 2006 to 2013 (Figure 2d). The y-axis refers to the number of uploaded videos. We start in 2006 with a very low interest in 3D videos, which increases until 2011 and then start decreasing until 2013. The initial sharp growth is due to the increase of popular 3D cinema movies, like Avatar which stimulated the public interest in 3D video content. At the same time stereo cameras became more accessible to the public as they cost less.

3. YOUTUBE 2D/3D VIDEO CLUSTERING

A semi-supervised [8] and a unsupervised [9] method was applied in the YouTube video graph. The semi-supervised method belongs to the family of label propagation techniques. Using a seed, i.e. a label, in a graph we let the algorithm spread this label to the nodes with similar characteristics based on the graph structure. We place these seeds, i.e. the initial labels, in five and ten nodes with the highest degree. In the unsupervised method clustering using the Normalized Cuts [9], a grouping of the videos in five and ten clusters was performed.

The title of a video describes the content of the video. We qualify the correctness of each cluster by looking whether the titles that belong to a cluster describe linguistically similar content or not. More specifically, using the video titles that belong to a cluster a word histogram is computed (language English). The thirteen most frequent words were used for the clustering evaluation. We pruned words like dates and conjunctions. In Table 1 and 2, the linguistically incoherent words are italicized. In some clusters there are only incoherent words (Table 1, Cluster 10, semi-supervised method). This means that these videos that belong to these clusters are not semantically coherent. Also the words of these clusters have very small frequency on the contrary to words that belong to more meaningful clusters like the Cluster 1 in Table 1 of the semi-supervised learning. For example, in Table 1, for the semi-supervised method, Clusters 1,2,3 and 6 are pure clusters while Clusters 5 and 9 are less semantically connected.

Moreover, in some Clusters, like Cluster 9 in Table 1 produced by the unsupervised method, beyond the few words that have a semantic meaning there are many incoherent words in italics. The words that are semantically coherent in these clusters are less than the incoherent words but they have higher frequency. Also, in these clusters all the words have relatively small frequency, which implies that we have small clusters. An other interesting cluster example is Cluster 8, in Table 1, of the semi-supervised method. In this cluster it seems that we have a collection, compilation, of videos as the keyword compilation has high frequency.

In Table 1 we can see the results of the semi-supervised and unsupervised method, when working with ten clusters. Generally, the unsupervised method gave better results, when working with ten and five clusters. We can observe that both semi-supervised and unsupervised algorithm have detected similar groups when working with ten clusters. For example, in Table 1 Clusters 1 and 5 of the semi-supervised method have similar semantic content with the Cluster 3 of the unsupervised algorithm. Also Clusters 4 and 10 of the semi-supervised and unsupervised method, respectively, are semantically connected. Generally, we have clusters referring to extreme sports like snowboard, surfing, motocross, sports on snow, sports relevant to free falls and clusters that are relevant to movies and trailers. For example, in Table 1 in the results from the semi-supervised method Cluster 2 refers exclusively to ski sport while Cluster 7 is relevant to workout.

In Table 2, we have the results of the unsupervised method for five clusters. We omit the corresponding results of the semi-supervised method because they were not satisfying. We can see that in Cluster 1 we have words related to movies and trailers similar to the Cluster 1 in the results with ten cluster of the unsupervised method. Also in Table 2 we observe that in Clusters 2 to 5 the labels are not so pure as in the case we study with ten clusters. For example, in the Cluster 5 of Table 2 we have words relevant to ski and surfing, while these meanings are separated in different clusters, when working with ten clusters in the semi-supervised method.

Table 1. Ten clusters for the unfiltered graph

Groups from semi-supervised [8] learning									
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
snowboard(65)	ski(108)	gymkhana(18)	surfing(35)	snowboard(4)	space(33)	crossfit(9)	compilation(133)	mussikkivideo(5)	*gopro(9)*
snowboarding(41)	skiing(77)	pastrana(16)	surf(33)	snowboarding(2)	wingsuit(30)	workout(9)	fail(69)	*virallinen(5)*	*hero(4)*
ski(29)	freestyle(28)	rally(14)	surfer(22)	breck(1)	landing(24)	body(4)	failarmy(36)	lyrics(4)	*park(2)*
mountain(23)	gopro(51)	drift(8)	wipeouts(20)	jump(1)	supercross(22)	fitness(3)	fails(66)	*epic(19)*	*summer(1)*
downhill(15)	freeski(28)	car(7)	wave(19)	skiing(1)	air(18)	strongest(3)	funny(53)	*people(4)*	*breck(1)*
bike(12)	freeride(15)	gear(7)	hawaii(16)	*powder(1)*	flying(16)	bodybuilding(2)	best(111)	*rich(4)*	*ollie*
freestyle(11)	salomon(14)	motocross(6)	blanchard(13)	*minishred(1)*	corliss(16)	piana(2)	bmx(54)	*high(4)*	*partly(1)*
winter(11)	skis(11)	racing(6)	pipiline(12)	*gopto(9)*	shuttle(16)	bodyweight(2)	bike(75)	*compilation(5)*	*minishered(1)*
mtb(10)	backcountry(10)	crash(6)	wipeouts(20)	*highlights(1)*	jump(14)	*people(5)*	*funny(53)*	*ultimate(4)*	*highlights(1)*
bull(65)	powder(44)	ride(5)	*camera(49)*	*park(2)*	*jeb(16)*	*worlds(2)*	*episode(31)*	*russia(4)*	*bowl(1)*
red(65)	svindal(14)	*episode(6)*	*girl(14)*	*private(1)*	*winter(15)*	*london(2)*	*top(45)*	*rich(4)*	*ryan(1)*
top(10)	*circus(13)*	*playground(5)*	*big(14)*	*ryan(1)*	*camera(15)*	*real(2)*	*world(36)*	*world(4)*	*nail(1)*
people(10)	*team(11)*	*awesome(5)*	*bbc(10)*	*trip(1)*	*bull(14)*	*rich(3)*	*win(31)*	*bass(3)*	*powder(1)*
Groups from unsupervised [9] learning									
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6	Cluster 7	Cluster 8	Cluster 9	Cluster 10
trailer(92)	bicycle(14)	ski(148)	landing(37)	compilation(125)	games(66)	gopro(152)	official(26)	video(10)	gopro(43)
video(59)	rear(8)	skiing(89)	airport(19)	fail(71)	street(49)	bike(94)	video(24)	workout(10)	shark(37)
movie(51)	gopro(8)	gopro(83)	maarten(19)	fails(66)	gopro(43)	bmx(53)	music(music)	crossfit(9)	surfing(35)
transformers(29)	cassete(7)	snowboard(72)	cockpit(14)	funny(47)	skateboarding(24)	downhill(50)	musiikkivideo(5)	body(4)	wingsuit(33)
jackass(24)	shimano(7)	powder(51)	takeoff(14)	failarmy(36)	skateboard(19)	motocross(44)	virallinen(5)	fitness(3)	surf(31)
batman(22)	derailleur(6)	avalance(46)	crash(13)	pranks(21)	bmx(18)	mountain(44)	karjalainen(3)	*epic(5)*	extreme(24)
official(22)	wheel(6)	snowboarding(46)	vulcan(13)	prank(20)	xgames(15)	ken(38)	fatboy(3)	*year(5)*	wave(22)
titanic(21)	bike(3)	freeski(35)	boeing(11)	*best(85)*	skate(11)	games(37)	live(3)	*official(7)*	surfer(21)
animation(18)	freewhell(3)	wallisch(25)	jet(11)	*awesome(33)*	*section(21)*	moto(37)	feat(6)	*russia(4)*	hawaii(20)
full(17)	*adjust(9)*	mountain(22)	747(14)	*epic(27)*	*summer(15)*	mtb(35)	bass(3)	*beach(3)*	wipeouts(20)
glasses(44)	*build(5)*	*freestyle(40)*	xh558(9)	*amazing(14)*	*gold(14)*	mtb(35)	lewis(3)	*ultimate(3)*	*extreme(24)*
3d(386)	*archery(3)*	*red(22)*	air(13)	*part(29)*	*final(15)*	race(31)	*party(3)*	*top(4)*	*camera(76)*
part(27)	*hub(3)*	*bull(22)*	*gun(10)*	*week(35)*	*city(11)*	supercross(29)	*bronson(16)*	*year(5)*	*hero(171)*

Table 2. Five clusters for the unfiltered graph

Groups from unsupervised learning [9] groups				
Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
trailer(92)	workout(10)	compilation(128)	bike	ski(153)
movie(51)	crossfit(9)	fail(71)	bmx(71)	skiing(93)
transformers(29)	body(4)	fails(67)	downhill(52)	snowboard(80)
batman(22)	music(14)	funny(53)	motocross(49)	powder(57)
official(22)	mussikkivideo(5)	failarmy(36)	race(36)	snowboarding(52)
titanic(21)	lyrics(4)	pranks(21)	mtb(35)	avalanche(49)
animation(18)	macklemore(4)	prank(21)	suppercross(32)	shark(37)
avatar(14)	prank(8)	landing(37)	skateboarting(26)	surfing(35)
film(14)	compilation(5)	air(14)	racing(24)	freeski(35)
new(13)	official(33)	airport(20)	wheels(23)	surfing(35)
dark(20)	action(16)	takeoff(14)	biking(23)	wingsuit(33)
moon(16)	*live(6)*	cockpit(cockpit)	crash(37)	jump(31)
supper(23)	*action(16)*	*best(89)*	*bull(98)*	winter(29)

4. CONCLUSIONS

In the paper, we study the YouTube 2D/3D video recommendation structure. Both the unfiltered graph and the filtered graph show statistical characteristics similar to previous works on social media. The clustering results were very encouraging for the unfiltered graph. In future work we will try to improve the clustering results particularly for the filtered 3D graph case. Also we will study not only the degree distribution but also the distribution of likes, dislikes, and the category of each video.

5. REFERENCES

[1] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and analysis of online social networks," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, 2007, pp. 29–42.

[2] L. A. Adamic, O. Buyukkokten, and E. Adar, "A social network caught in the web," *First Monday*, vol. 8, June 2003.

[3] S. Milgram, "The small world problem," *Psychology Today*, vol. 32, pp. 425–443, 1967.

[4] Y.-Y. Ahn, S. Han, H. Kwak, Y.-H. Eom, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *In Proceedings of the 16th international conference on World Wide Web (WWW)07*, 2007, pp. 835–844.

[5] S. Wasserman and K. Faust, "Social network analysis: Methods and applications," 1994.

[6] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, January 2006.

[7] L. Becchetti, C. Castillo, D. Donato, A. Fazzone, and I. Rome, "A comparison of sampling techniques for web graph characterization," in *Proceedings of the Workshop on Link Analysis (LinkKDD06), Philadelphia, PA*, 2006.

[8] D. Zhou, O. Bousquet, T. Navin Lal, J. Weston, and B. Schlkopf, "Learning with local and global consistency," in *Advances in Neural Information Processing Systems 16*. 2004, pp. 321–328, MIT Press.

[9] J.Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888–905, 1997.