A Correlation Analysis of Web Social Media

Konstantinos N. Vavliakis
Dept. of Electrical and Computer Engineering
Aristotle Univ. of Thessaloniki
&
Informatics and Telematics Institute, CERTH
Thessaloniki, Greece
kvavliak@issel.ee.auth.gr

Konstantina Gemenetzi
Dept. of Electrical and Computer Engineering
Aristotle Univ. of Thessaloniki
Thessaloniki, Greece
kgemenet@auth.gr

Pericles A. Mitkas
Dept. of Electrical and Computer Engineering
Aristotle Univ. of Thessaloniki
&
Informatics and Telematics Institute, CERTH
Thessaloniki, Greece
mitkas@eng.auth.gr

ABSTRACT

In this paper we analyze and compare three popular content creation and sharing websites, namely Panoramio, YouTube and Epinions. This analysis aims in advancing our understanding of Web Social Media and their impact, and may be useful in creating feedback mechanisms for increasing user participation and sharing. For each of the three websites, we select five fundamental factors appearing in all content centered Web Social Media and we use regression analysis to calculate their correlation. We present findings of statistically important correlations among these key factors and we rank the discovered correlations according to the degree of their influence. Furthermore, we perform analysis of variance in distinct subgroups of the collected data and we discuss differences found in the characteristics of these subgroups and how these differences may affect correlation results.

Although we acknowledge that correlation does not imply causality, the discovered correlations may be a first step towards discovering causality laws behind content contribution, commenting and the formulation of friendship relations. These causality laws are useful for boosting the user participation in social media.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences - Sociology; G.3 [**Probability and Statistics**]: Correlation and regression analysis

General Terms

Measurement, Human Factors

Keywords

Social Media,Regression Analysis, ANOVA, Correlation

WIMS '11, May 25-27, 2011 Sogndal, Norway

1. INTRODUCTION

A key factor for the success of the so-called Web Social Media (WSM) is that they allow users to form virtual communities, actively participate in the content creation and publication process, as well as leave feedback and express opinions in the form of comments. In fact WSM success heavily depends on mass user participation. Nowadays any website with obsolete information instantly ceases to exist. As a result, WSM owners are relieved from the labor of updating their site with new content and concentrate their efforts on optimizing the means for delivering content together with motivating users to keep contributing new content.

An in-depth understanding of the graph structure and statistical properties of WSM is necessary to better understand existing websites and the communities formed within them, evaluate their impact, improve their interfaces and design new network-based web systems. Although many efforts have focused on the analysis of WSM, there is still a great level of obscurity in completely understanding their properties.

In our study we concentrated on three popular WSM with different functionalities, Panoramio[1], YouTube[2] and Epinions[3]. All of them bear the same basic characteristic of contributing and sharing online sources (photo, video or text input). Panoramio is a geolocation-oriented photo sharing website that can be accessed as a layer in Google Earth and Google Maps. YouTube, is the largest video related website on which users can upload, share and view videos. In November 2010 it has been reported that 35 hours of video is uploaded in YouTube every minute [13]. Finally, Epinions is a general consumer review site. At Epinions, visitors can read reviews about a variety of items or they can write reviews in exchange of money and recognition. Epinions pays Income Share as reward to reviewers helping other users to purchase products. This feedback mechanism along with Epinions' reputation system enroll Epinions into a special WSM category and distinct some of its characteristics. Although the selected WSM exhibit differences in the media type they host, as well as in their operation scheme, they all constitute virtual communities, in which users have to sign in and can upload files, view and comment on other users' files.

The focus of this work is to compare statistical proper-

[1]http://www.panoramio.com/
[2]http://www.youtube.com/
[3]http://www.epinions.com/

ties and correlations of the three popular WSM. We analyze their common characteristics and calculate correlations of their attributes using regression analysis. We also compare subsamples of WSM available data and test the variance of their statistical properties in order to discover whether their correlations are caused by other factors. Moreover although correlation does not imply causation, our correlation results may be a strong indication for further testing on causal relationships. Discovery of causal relationships in WSM is necessary in order to determine the motives of user participation and stimulate users into increasing their participation, thus increasing WSM's revenue.

The paper is organized as follows: In Section 2 we present related work and motivation, while in Section 3 we introduce the data collection mechanisms and present statistical properties of the collected data from three different popular Social Media. Section 4 contains the correlation analysis results, and the subgroup analysis of variances. Finally, Section 5 discusses conclusions and proposes future work directions.

2. RELATED WORK

The social networks forged within WSM can be used in numerous information extraction tasks, like expert identification, backbone discovery for tracking influential individuals, and structural analysis of user interaction that offers insight into the macroscopic behavior of online social networks. Moreover, regression analysis and analysis of variance in discovering motivational factors that affect content contribution and sharing in WSM have been used in [9]. Probabilistic models in showing that interacting users are much more likely to share many similarities than any pair of random users have been used in [12] .

Many studies analyze YouTube data. A systematic and in-depth measurement study, which compares YouTube videos to traditional streaming videos is presented in [4], while [2] offers a large scale analysis of YouTube in the frame of user generated content systems, in which information such as the life-cycle of videos and the intrinsic statistical properties of requests are presented. On the other hand, [10] studies the structure and network characteristics of YouTube, and shows that relationships such as friendship and commenting are shaping YouTube interaction.

Considerably fewer studies are focused on Panoramio, nevertheless, there are some interesting works, like [1], which uses the space and time referenced photos of Panoramio to discover interesting places, temporal patterns of visits to places and to investigate flows between places. Flickr[4], a WSM bearing many similarities with Panoramio, has been the focus of several studies. Reference [8] focuses on the ways new links are formed and investigates the link formation process, while [3] studies social cascades, that is how information disseminates through social links in Flickr. Moreover, [6] shows that social browsing is one of the primary methods by which users find new images in Flickr.

Finally [5] presents a framework for learning link prediction on Epinions and other web sources while [7] provides insight into fundamental principles that drive the formation of signed links in Epinions.

Although very interesting most of the works are performed in a narrow scope, mainly due to the lack of easy-to-use/build publicly available datasets, enriched with suitable metadata. Additionally most of the efforts analyze a single WSM, while in case numerous WSM are examined, the analysis is usually narrowed in comparing their statistical properties and network structure.

[4]http://www.flickr.com/

3. CRAWLING AND DATA EXPLORATION

The first step in our data collection process involved crawling and HTML scraping the selected websites. The collected data of the three WSM were organized and homogenized into a common database schema. In the preprocessing phase the data logarithms were calculated (due to power-laws). Finally, regression analysis and analysis of variance were performed, in order to discover interesting correlations among users, uploaded files, feedback received in the form of comments or views, and tenure.

The adopted methodology aspires to infer knowledge about user online activities, as well as correlations among these activities. The results of this analysis can provide WSM useful insight and the necessary means to understand feedback mechanisms, correlation phenomena and ultimately increase their traffic and content, thus increase their social and financial capital.

For our study needs, we collected 1.500 *complete* profiles of Panoramio and Epinions users and 10.000 profiles of YouTube users. A complete profile includes all the available information of a single user. That is his/her online available personal information, all files (photos, videos or reviews) the user has posted, as well as comments and tags his/her photos received. The statistical properties of the most important attributes of the collected data are shown in Table 1. As expected, most numerical attributes follow a power law distribution, a usual phenomenon in cases of data derived from social networks [11]. In this case, logarithmic values were used for linearizing data. Figure 1 depicts the relations of users with the logarithms of various factors.

The collection process was automatic, through three independent crawlers, each one assigned to collect data from a different WSM. Although all crawlers share a common collecting and indexing mechanism, the different HTML structure of the examined WSM forced the development of three self-contained crawlers. Nevertheless, all crawlers share the same database system for storing crawled information.

The developed crawlers, first collect a random set of files. Then a bundle of information for each user appearing in the initial set is gathered. To this end we avoid falling into a dense cluster of users with similar properties and we acquire a randomly distributed set of users. For each crawled user, the following information is retrieved: profile info, files he/she has posted, comments and tags these files have received, groups they participate, favorite files and first and second degree contacts. All user information is anonymously processed. The crawling mechanism keeps track of the profiles visited, enabling stop and restart, without having to revisit already crawled profiles. It should also be noted that the crawling algorithm implementation is distributed, allowing for the parallel execution of the algorithm.

Finally, we used the $|t|$ variable for ranking variables according to their influence level to the dependent variable. According to the results of Table 2, we ranked the correlation degree of the independent variables with the dependent one. The ranking (in descending order) of the most influential variables for each WSM is depicted in Table 3.

Attributes		Panoramio	YouTube	Epinions
Users	Total	1.500	10.000	1.500
Files	Total	194.871	184.833	31.229
Views per file	Min/Max	2 / 461.510	0 / 49.392.222	-
	Mean / Std. Dev:	386,83 / 2.161,97	8379,08 / 190.142,05	-
Total Views		75.382.765	1.548.729.945	-
Comments per file	Min/Max	0 / 2.249	0 / 114.447	0 / 704
	Mean / Std. Dev:	1,82 / 11,36	21,56 / 411,79	2,41 / 7,93
Comments	Total	355.242	3.931.911	74.554
Tags per file	Min/Max	0 / 50	0 / 89	-
	Mean / Std. Dev:	3,12 / 3,50	9,22 / 9,12	-
Tags	Total	607.430	1.704.308	-
Files per user	Min/Max	0 / 6.280	0 / 3.277	0 / 1.090
	Mean / Std. Dev:	198,20 / 400,51	18,57 / 84,75	20,74 / 59,60
Favorite users per user	Min/Max	0 / 1,14	0 / 11,47	0 / 1,57
	Mean / Std. Dev:	10,01 / 43,36	27,30 / 168,55	12,94 / 58,95
Trusting users	Min/Max	-	0 / 163.892	0 / 1.174
	Mean / Std. Dev:	-	133,24 / 2467,78	12,19 / 57,08
Favorite files per user	Min/Max	0 / 2.904	0 / 1.000	-
	Mean / Std. Dev:	2,74 / 40,62	96,58 / 176,38	-

Table 1: Statistical properties of collected data.

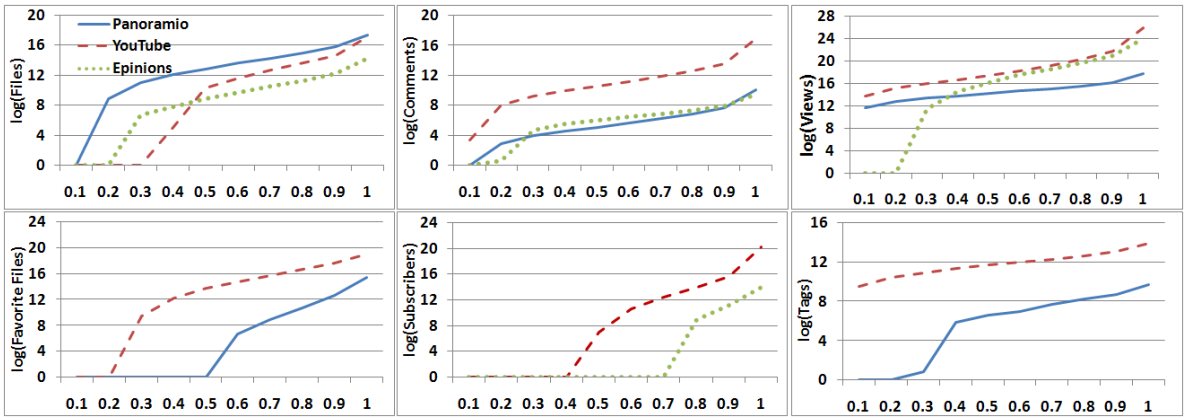

Figure 1: Various factors in the examined WSM. X axis denotes percentage of the users, ranked by the logarithmic attribute value depicted in Y axis.

4. CORRELATION ANALYSIS

4.1 Regression Analysis

For identifying correlations among various factors of WSM, we selected similar attributes appearing in the crawled WSM. These attributes are: *a)*Files: the number of files uploaded by users, that is photos in case of Panoramio, videos in case of YouTube, and product reviews in case of Epinions, *b)*Tenure: this is the time period a user is a member of the WSM community, or in case this information is not available, the elapsed time since the first file upload, *c)*Views: which is the number of views per file in the case of Panoramio and YouTube, while, in the case of Epinions, total visits, as calculated by the Epinions' algorithm, *d)*Friends: which are users participating in user's personal social network, and *e)*Comments: representing the number of comments received per uploaded file.

We used regression analysis to extract correlation among these attributes. We conducted three experiments using the regression model $Y \approx f(X, \beta)$, by setting as the dependent variable Y the $\log(comments)$, $\log(views)$ and $\log(files)$ respectively and as independent variables $X_i, (1 \leqslant i \leqslant 4)$ the remaining four attributes. Detailed results of the analyses are depicted in Table 2. in Variable β denotes the coefficient or in other words the least square estimate, which shows whether there is a positive or negative correlation between two variables. Variable t is the coefficient divided by the standard error (computed t-statistic for test of H0, also known as Null hypothesis) and denotes how important is the coefficient in the overall model. Finally p is the p-value of H0-Null Hypothesis against Ha, or in other words the confidence interval. In our analysis we used 95% confidence intervals, that is p-values smaller than 0,05.

Table 2 also provides general information about regression analysis, such as R^2, Adjusted R^2 and F. R^2 denotes the correlation between Y (the true value of the dependent variable, number of comments views and file in our case) and Y-hat (value of the dependent variable predicted), Adjusted

	Panoramio			YouTube			Epinions		
	β	t	p	β	t	p	β	t	p
	Comments								
Friends	0,067	4,609	0,000	0,051	7,311	0,003	-0,018	-0,786	0,432
Tenure	0,270	0,830	0,407	-0,309	-4,386	0,000	-1,775	-4,568	0,000
Views	0,993	7,686	0,000	1,182	39,855	0,000	0,444	9,672	0,000
Files	0,454	5,989	0,000	0,672	15,907	0,000	1,393	8,908	0,000
	R^2=0,174, F = 52,339			R^2=0,350, F = 710,249			R^2=0,286, F = 133,323		
	Views								
Friends	0,015	4,211	0,000	0,024	8,488	0,000	0,003	0,246	0,806
Tenure	1,474	23,828	0,000	0,736	27,437	0,000	1,997	8,438	0,000
Comments	0,056	7,686	0,000	0,195	39,855	0,000	0,172	9,672	0,000
Files	-0,129	-7,175	0,000	0,097	5,538	0,000	1,263	13,524	0,000
	R^2=0,435, F = 191,931			R^2=0,143, F = 220,781			R^2=0,418, F = 202,318		
	Files								
Friends	0,073	13,209	0,000	0,023	10,387	0,000	0,074	20,899	0,000
Tenure	0,689	5,220	0,000	0,088	3,929	0,000	-0,113	-1,562	0,118
Views	-0,382	-7,175	0,000	0,059	5,538	0,000	0,110	13,524	0,000
Comments	0,077	5,989	0,000	0,068	15,907	0,000	0,047	8,908	0,000
	R^2=0,224, F = 71,766			R^2=0,402, F = 889,998			R^2=0,562, F = 361,829		

Table 2: Regression analysis results.

R^2 is equal to $1-(1-R^2)*(n-1)/(n-k-1)$ and F is the F-test value. In all equations n is the number of examined instances (degrees of freedom) and k is the number of regressors.

As depicted in Table 2 most of the attributes are not only highly correlated, but correlations are statistically important too (p-values smaller than 0,05), revealing the strong relations among similar fundamental factors in various WSM. As far as Panoramio and YouTube are concerned, the only insignificant correlation is between tenure and comments in case of Panoramio. On the other hand, Epinions has both significant and insignificant correlations. According to our analysis, the number of comments, as well as the number of views per file/review one receives are insignificantly correlated to the number of friends the user has. This is probably due to the fact that most comments in Epinions are of consultive nature, rather than of encouraging nature, so experienced users with many friendship relationships may receive limited number of comments. As far as the absence of correlation between number of friends and number of views, this is probably due to the custom view counter algorithm Epinions uses, and by which the Income Share reward is distributed. The number of reviews and tenure are also uncorrelated. This could be explained by the fact that many users are inactive most of the time, thus veteran users not necessary upload/write more reviews. All these non-correlations may also be due to the Income Share reward and reputation system, used by Epinions.

Finally, we used the $|t|$ variable for ranking variables according to their influence level to the dependent variable. We ranked the correlation degree of the independent variables with the dependent one according to Table 2. The ranking (in descending order) of the most influential variables for each WSM is depicted in Table 3.

Dep. Var.	Independent Variable
Panoramio	
Comments	1.Views 2.Files 3.Friends
View	1.Tenure 2.Comments 3.Files 4.Friends
File	1.Friends 2.Views 3.Comments 4.Tenure
YouTube	
Comments	1.View 2.Files 3.Friends 4.Tenure
Views	1.Comments 2.Tenure 3.Friends 4.Files
Files	1.Comments 2.Friends 3.Views, 4.Tenure
Epinions	
Comments	1.Views 2.Files 3.Tenure
Views	1.Files 2.Comments 3.Tenure
Files	1.Friends 2.Views 3.Comments 4.Tenure

Table 3: Ranking of the most influential variables.

4.2 Analysis of Variance

Further elaborating on the correlation results, we conducted an analysis of variance (ANOVA). We divided collected users of each WSM into three equal sized subsamples (low, medium and high) according to values of the dependent variables, examined in Section 4.1 (Comments, Views and Files). Then, we compared the average values of the remaining four independent variables of these subsamples, looking for significant differences. Results are depicted in Table 4, where the attributes in which difference greater than 50% among subsamples was observed are indicated.

In case of Panoramio, large variations were found in the total number of friends and the number of views per files/photo, in the subsamples of different number of comments. On the other hand, the subsamples of different number of comments had similar values regarding tenure and number of files/photos. Likewise, large variances were observed in the number of uploaded files/photos in the subsamples regarding different number of views and in the number of total friends in the subsamples of different number of files/photos. More results are available in Table 4, which presents ANOVA results for YouTube and Epinions as well. Epinions, once again, exhibits the greatest number of factors having large variations in different subsamples.

It is also notable that several subgroups exhibit similar

	Panoramio	YouTube	Epinions
	Comments		
Friends	√	√	√
Tenure			
Views	√	√	√
Files			√
	Views		
Friends		√	√
Tenure			
Comments		√	√
Files	√	√	√
	Files		
Friends	√	√	√
Tenure			
Views			√
Comments			

Table 4: Analysis of variance results.

mean values of certain attributes. For example, the mean value of tenure is similar in all subgroups of the examined WSM. This fact poses interesting questions on *how* tenure is correlated with the other variables, a topic for future research. Although preliminary, this analysis in combination with the results of correlation analysis, may be a first step towards discovering causality laws in WSM.

5. CONCLUSIONS AND FUTURE WORK

In this work we studied similar features appearing in three popular WSM, namely Files (in the form of photos, videos and reviews), Friends, Comments, Views and Tenure. After comparing the different statistical properties of these WSM, we performed regression analysis in order to discover and rank statistically important correlations between attributes. Although some results may have intuitive explanations, there is strong indication that there are strong correlations between many attributes in the examined WSM. The WSM with the least correlation effects was found to be Epinions, a fact that we attribute to the special nature of this WSM (income reward and reputation system).

We presented several ANOVA experiments and tested the variances when different subgroups are divided according to various factors. Several subgroups were found to have great differences in their statistical attributes, an effect requiring further testing and experimentation for more concise conclusions. Nevertheless, one must be careful in extracting firm conclusions, as different samples, or samples taken in different time periods, may lead to different results. Despite the inconsistencies in different subgroups, the correlations discovered may be a first step towards extracting causality of different attribute results, thus creating feedback and rewarding mechanism for increasing user content sharing and participation in WSM. This is a crucial task for WSM as their main income source is through targeted advertisements and is proportional to the size of their community and the average time users spend online.

Further testing and experimentation on larger datasets of more WSM is within our plans, for validating the correlations we discovered. Moreover, a natural extension of this work would be the investigation of causality in WSM by applying correlation analysis in controlled subgroups with similar properties. This could lead to the creation of feedback mechanisms for rewarding high-valued users.

6. REFERENCES

[1] G. Andrienko, N. Andrienko, P. Bak, S. Kisilevich, and D. Keim. Analysis of community-contributed space-and time-referenced data (example of panoramio photos). In *GIS '09: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 540–541, New York, NY, USA, 2009. ACM.

[2] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, New York, NY, USA, 2007. ACM.

[3] M. Cha, A. Mislove, B. Adams, and K. P. Gummadi. Characterizing social cascades in flickr. In *WOSP '08: Proc. of the first workshop on Online social networks*, pages 13–18, New York, NY, USA, 2008. ACM.

[4] X. Cheng, C. Dale, and J. Liu. Statistics and social network of youtube videos. In *Quality of Service, 2008. IWQoS 2008. 16th International Workshop on*, pages 229–238, 2008.

[5] J. Kunegis and A. Lommatzsch. Learning spectral graph transformations for link prediction. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, pages 561–568, New York, NY, USA, 2009. ACM.

[6] K. Lerman and L. Jones. Social browsing on flickr. In *International Conference on Weblogs and Social Media (ICWSM 2007)*, Boulder, Colorado, March 2007.

[7] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 641–650, New York, NY, USA, 2010. ACM.

[8] A. Mislove, H. S. Koppula, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Growth of the flickr social network. In *WOSP '08: Proc. of the first workshop on Online social networks*, pages 25–30, New York, NY, USA, 2008. ACM.

[9] O. Nov, M. Naaman, and C. Ye. Motivational, structural and tenure factors that impact online community photo sharing. In *International AAAI Conference on Weblogs and Social Media*, 2009.

[10] J. C. Paolillo. Structure and network in the youtube core. *Hawaii International Conference on System Sciences*, 0:156, 2008.

[11] J. P. Scott. *Social Network Analysis*. Sage Publications Ltd, London, UK, 2000.

[12] P. Singla and M. Richardson. Yes, there is a correlation: - from social networks to personal behavior on the web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 655–664, New York, NY, USA, 2008. ACM.

[13] YouTube. Youtube fact sheet, http://youtube-global.blogspot.com/2010/11/great-scott-over-35-hours-of-video.html, November 2010.