Towards Understanding How Personality, Motivation and Events Trigger Web User Activity

Konstantinos N. Vavliakis
Aristotle University of Thessaloniki
Electrical and Computer Eng. Dept.
Thessaloniki, Greece
+302310996349
kvavliak@issel.ee.auth.gr

Andreas L. Symeonidis
Aristotle University of Thessaloniki
Electrical and Computer Eng. Dept.
Thessaloniki, Greece
+302310994344
asymeon@eng.auth.gr

Pericles A. Mitkas
Aristotle University of Thessaloniki
Electrical and Computer Eng. Dept.
Thessaloniki, Greece
+302310996390
mitkas@eng.auth.gr

ABSTRACT
Moving on from Web 1.0, where information was static, Web 2.0 has provided internet users with a dynamic media, where information is updated continuously and anyone can participate. This new form of communication is enjoying impressive resonance, and has triggered the development of the so-called Social Media, which are fast growing enterprises profiting either by users' subscriptions, web advertisements or service exploitation. In any case, their flourishing and profitability are based on users' interaction, online habits and updated content. Though preliminary analysis exists, there is still little understanding on what exactly stimulates users to actively create and share their content online.

Within the context of this work we propose a methodology that aspires to identify and analyze those events that trigger web user activity, content creation and sharing in Web 2.0. This way, Social Media owners are expected to better understand the creation process of their main commodity. The methodology is based on user personality and motivation, and on the occurrence of events with a personal, local or global impact. The proposed methodology was applied on data collected from Flickr and analysis was performed through the use of statistics and data mining techniques.

The correlation of specific events with increased web user activity was identified. The impact of community feedback was also evaluated. Classification was performed in order to categorize users with respect to the number of images they have uploaded and with respect to the active time they have spend uploading photos. Association rule extraction was performed in order to discover non-trivial rules on web user activity.

Categories and Subject Descriptors
J.4 [**Computer Applications**]: Social and Behavioral Sciences – Sociology; H2.8 [**Database Applications**]: Data Mining

General Terms
Measurement, Experimentation, Human Factors

Keywords
Crowdsourcing, Data Mining, Flickr, Sharing, Social Media

1. INTRODUCTION
During the last few years we have witnessed a radical transformation of the WWW, from a read-only infrastructure to a read-write-debate corpus. In this so-called Web 2.0 [22], millions of webpages, blogs, forums etc are updated on an hourly (sometimes near real-time) basis, providing anyone with an internet connection the opportunity to state his/her opinion or experience on practically any given subject, upload interesting or even personal images and videos and connect to friends and interesting people. Having enjoyed the 24/7 availability of information on the Web, users have now started to develop rigorous web "social lives". They actively participate in content creation in the form of news, videos, music, blog posts, etc, providing information on their political and religious views, their likes and dislikes, their dreams and fears. In order to so, they use popular **Web Social Media** (WSM), like Facebook, Wikipedia, YouTube and Flickr.

WSM generate a deluge of chaotic information, and in turn, substantial opportunities to profit from the newly-formed online societies. WSM success heavily depends on mass user participation; a WSM with obsolete information instantly ceases to exist. WSM owners are relieved from the labor of updating their site with new content and are now concentrated in optimizing the means of delivering content together with motivating users to keep contributing new content.

This work aims to explain how various factors affect user activity in the web, thus providing the necessary means to WSM owners to increase user participation and contribution, improve their social capital and ultimately upgrade and expand their corporate models.

Preliminary analysis has shown that the pivotal factors for participation are not of financial nature, rather related to social psychology. The goals of this study are to identify: *1)* user personality characteristics that urge them to participate and contribute content in online communities, *2)* motivational factors and their relation to personality *3)* the way feedback received from the community affects participation and *4)* events occurrence

(with a personal, small-scale or large-scale impact) that drive users to change their online behavior.

The paper is organized as follows: In Section 2 we present related work and motivation, while in Section 3 we introduce the proposed model and the methodology of our study. Section 4 discusses data collection and further elaborates on the analysis conducted. In particular, Section 4.1 presents the necessary preprocessing layer implemented, while Section 4.2 analyzes how motivation and community feedback affect users. Section 4.3 deals with the effect of world and personal events on the sharing ratio of photos. Consequently, Section 4.4 discusses the estimation models built for classifying users according to their "web behavior", while Section 4.5 focuses on some interesting observations made, based on association rule extraction analysis. Finally, Section 5 discusses conclusions and proposes future work directions.

2. RELATED WORK & MOTIVATION

Numerous approaches exist attempting to identify the key factors that trigger user activation in Web 2.0. It is a task that involves multidisciplinary research, since one has to combine theories developed in the psychological domain, along with social and computer science advances, in order to come up with a viable interpretation of web user behavior.

From a psychology perspective, personality traits and individual differences clearly affect one's web activity. Literature is rich on works analyzing people and personalities ([2],[3],[17]). One would find analysis in [2] particularly intriguing, proposing two ways of studying personality: the *nomothetic* psychology, which seeks general laws to apply to groups of different people, and the *idiographic* psychology, which tries to grasp unary characteristics of people. In current work we adopt the nomothetic psychology dogma, trying to elicit generalized rules on web activity.

From a social science perspective, behavior is affected by the presence of others ([4],[10],[18],[27]). In this context, six motivational categories are identified [20], namely: a) *Values* (related to altruistic and humanitarian concerns), b) *Social* (the chance to be with friends or to engage in activities viewed favorably by important others), c) *Understanding* (learn new things and exercise one's knowledge), d) *Career* (an opportunity to achieve job-related benefits), e) *Protective* (reducing guilt over being more fortunate than others, or addressing one's own personal problems), and f) *Enhancement* (serve the ego and publicly exhibit one's knowledge). On the other hand, Wasko and Faraj [29] argue that two distinct sets of factors may be salient in the context of content sharing: a) *individual motivations*, like enjoyment, commitment to the community, self-development and reputation attainment and, b) *network structure factors*, like the degree users are embedded in a community. Finally, Lambel and Bhalla [14] introduce *status* and *recognition* as key motivators for contributing in online communities.

Based on the aforementioned qualitative findings, recently published papers present interesting quantitative results. Non et al. [21] integrate motivational factors measured via online surveys with structural properties of users' profiles available online, in order to understand user motivation in contributing to photo-sharing online communities. Though promising, this work focuses only on the current status of each user, failing to address questions on the way user behavior evolves in time and the triggers for his/her actions within the time course. On the other hand, Huberman et al. [12] attempt to explore user behavior in the course of time, nevertheless shrinking their criteria basis only on the attention (measured as the number of views) as a factor affecting users' participation. Other interesting papers that have analyzed the characteristics and user of web and its social media include [26], where the usage of web is studied, and [6], where users' web activity is evaluated in order to enhance advertising effectiveness. Moreover, online communities and content sharing websites have been widely studied in terms of statistical characteristics [7], browsing habits [16] and the development of special cultures and practices [19], as well as in terms of knowledge sharing [30] and diffusion [5].

Nevertheless, none of the abovementioned approaches succeeds in forming the bigger picture of the Social Web. Research is performed in a narrow scope, mainly due to the lack of easy-to-use/build publicly available datasets, enriched with metadata. Additionally, most of the work performed does not take the evolution of web user behavior over the course of time into account.

Further expanding works of Non et al. [20] and Huberman et al. [12], we have built a generalized model and the respective methodology for combining events occurrence, personality and motivations. We map numerous publicly available features regarding users' profile into this model in an attempt to understand how these features affect online activities, namely content contribution. Moreover, through the analysis of users' behavior over time, we aspire to identify how users adapt their online behavior according to external stimulus and feedback of the online community.

3. MODEL & METHODOLOGY

People surf the web for millions of different reasons. Some of them may be due to some personal, local or global event. Even the announcement of a new web site or an update in an existing web service may trigger a whole new series of user online activities. The repercussion of these events in one's web activity mainly depends on his/her personality, which by itself is a discernible factor greatly affecting (re)action to web stimuli. One should keep in mind, though, that personality and motivation are strongly interconnected, since motivation also forms personality.

To this end, we consider three pivotal factors affecting online activities, namely: 1) *events*, 2) *personality* and 3) *motivations*. As depicted in Figure 1, events affect personality, on which motivations are dependent. The aggregation of these three fundamental factors defines web user activity, which in turn affects users' motivational factors through feedback received via their online activities.

We have attempted to formalize the three primal factors affecting online activities by mapping them to freely available online data. We have chosen a popular photo sharing website as a testbed for our analysis. Flickr[1] is an image hosting website and an online community currently owned by Yahoo Inc. Flickr has also added video hosting services. According to latest reports [8], Flickr hosts more than 4 billion images. Taking into account that an average user has posted around 2.000 images [21], this means that approximately 4 million people are using (or have used) the Flickr web services.

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

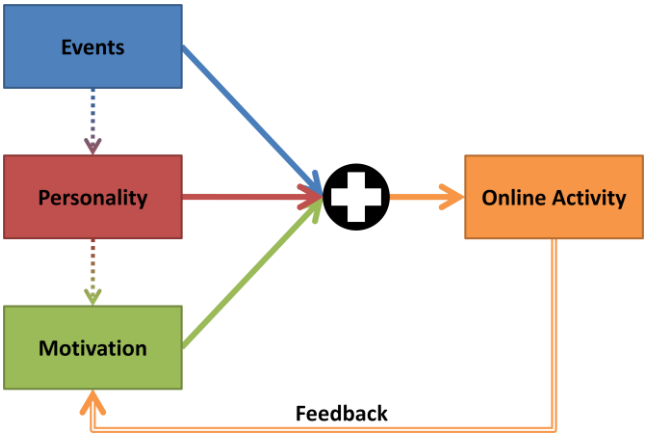

Figure 1: Factors affecting activity in online communities

Figure 2 depicts the 4-step methodology adopted. First, we collected data by crawling the social network of Flickr. Then, data are normalized and discretized in the preprocessing phase. Statistical and data mining analysis is then performed, in order to discover interesting patterns regarding user participation in the online community. This way, we expect to infer knowledge about users' online activity that can provide WSM with the necessary means to increase their social capital and their content creation (and upload) rates.

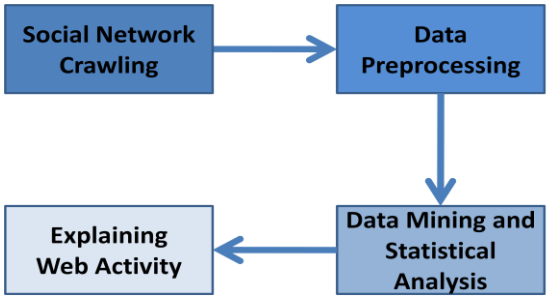

Figure 2: Methodology Steps

4. DATA ANALYSIS AND RESULTS

Data used for analysis are Flickr data, gathered online from November 5th to December 30th 2009. The collection process was automatic, through the developed crawler, that employs a depth-first approach built on flickrj [9], a free Java API which wraps the REST based Flickr API.

The crawler designed first seeks for geotagged photos taken in various capitals of the world and then a bundle of information of every user that has uploaded a photo in one of these capitals is collected. For each user, the following information is retrieved: profile info, the photos they have posted, the comments and tags these photos have received, the groups they participate, their favorite photos and their 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. Moreover, the crawling algorithm implementation is distributed, allowing for the parallel execution of the algorithm.

By the use of the crawler we collected data on 741 users. That is, 922.487 public photos, which received 1.558.822 comments and were tagged 7.987.827 times. Preliminary analysis showed that out of the 741 users, 448 (60.46%) were male, 108 (14.57%) female and 185 (24.97%) unknown. The most interesting user features are summarized in Table 1, where a short description of each feature and some statistical info are provided. It should be mentioned at this point that, like in all cases of data derived from social networks [27], all numerical attributes follow a power law distribution.

Figure 3 depicts the fluctuation of photos in our dataset with respect to time. The horizontal axis represents the time scaled in weeks, starting from Week 1 which stands for January 1st 2000. The continuous line denotes the date photos were uploaded in Flickr, while the dashed one determines the date they were taken, according to their Exif [28] metadata information.

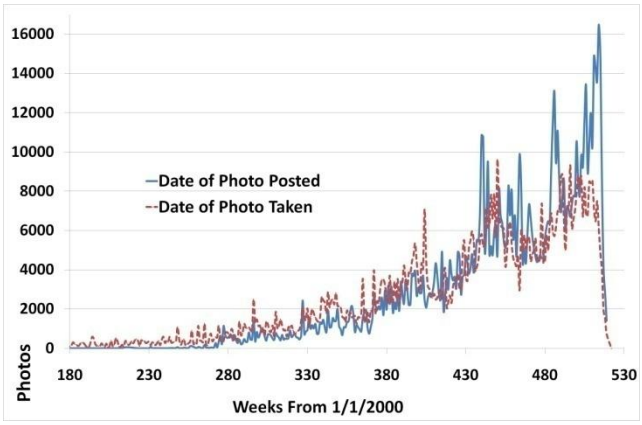

Figure 3: Fluctuation of photos, according to date they were taken and the date they were posted online

Next, the preprocessing phase is discussed, along with the analysis conducted, The findings, apart from a better understanding of how the Flickr online community interacts, offer valuable insight on how one could increase user participation and contribution to online communities, grow WSM social capital and, boost content transactions among users.

4.1 Data Preprocessing

The data collected were preprocessed in order to reduce noise and erroneous instances. No outliers were removed, as they were expected due to the power law distributions existing in social network data. During the preprocessing phase, data were cleaned as there were cases of erroneous data (e.g. photos with future dates in the Exif date taken field), the appropriate data categories for building estimation models were selected and the numerical attributes were discretized into three equal bins for dealing with ordinal data, quantifying this way each attribute into three classes: '*Low*', '*Medium*' and '*High*'. Due to the power law distribution, discretization was performed with respect to equal instance frequency in each bin, rather than equal discretization ranges. For performing meaningful time series analysis, data series were also normalized, partitioned into one week periods and chronologically deployed, setting the start of the analysis at Week 1 - January 1st 2000 (discussed above).

Table 1. Users' information collected from Flickr

	Information	Description	Statistical Properties
1	Is pro	Users having a pro account. The main difference between free and pro accounts is that free accounts display only the 200 most recent images.	**Pro**: 530 (71.52%) **Non pro**: 211 (28.48%)
2	Sex	Users' sex according to their profile.	**Male**: 448 (60.46%) **Female**: 108 (14.57%) **Null Values**: 185 (24.97%)
3	Status	Users' status according to their profile.	**Taken**: 234 (31.58%) **Single**: 80 (10.80%) **Null Values**: 427 (57.62%)
4	Testimonials	Users can leave a testimonial to another user, usually extolling their photograph techniques.	**Total**: 408 (in the dataset) **Min**: 0, **Max**: 33 (per user) **Mean**: 0,55 **StdDev**: 2,24
5	Participation Time	Total time (in weeks) between user's first photo upload and 31/12/2009.	**Total**: 517.085 (in the dataset) **Min**: 0, **Max**: 14.563 (per user) **Mean**: 698,76 **StdDev**: 732,52
6	Active participation Time	Total number of weeks in which users have uploaded at least one photo.	**Total**: 34.980 (in the dataset) **Min**: 0, **Max**: 263 (per user) **Mean**: 47,27 **StdDev**: 45,62
7	Group	Groups in which users participate. Usually groups' purpose is to collect different users' images regarding a specific theme.	**Total**: 70.509 (in the dataset) **Min**: 0, **Max**: 2058 (per user) **Mean**: 95,15 **StdDev**: 154,18
8	Contacts	Number of users' contacts.	**Total**: 93.234 (in the dataset) **Min**: 1, **Max**: 1710 (per user) **Mean**: 125,82 **StdDev**: 207,83
9	Favorite Photos	Selected images that are the favorites ones for each user.	**Total**: 461.276 (in the dataset) **Min**: 0, **Max**: 21.865 (per user) **Mean**: 622,5 **StdDev**: 1809,53
10	Comments per photo	Comments received per uploaded photo.	**Total**: 1.558.822 (in the dataset) **Min**: 0, **Max**: 158,97 (per user) **Mean**: 4,27 **StdDev**: 11,77
11	Tags per photo	The average number of tags in each photo.	**Total**: 7.987.827 (in the dataset) **Min**: 0,024, **Max**: 39,18 (per user) **Mean**: 7,53 **StdDev**: 6,01
12	Geolocated Photos	Photos geotagged with longitude and latitude information.	**Total**: 548.267 (in the dataset) **Min**: 1, **Max**: 29.697 (per user) **Mean**: 733,9 **StdDev**: 2205,75
13	Travelled Distance	Total distance (in KMs) travelled by users as estimated by their chronologically ordered geolocated photos.	**Total**: 51.861.407 (in the dataset) **Min**: 0, **Max**: 2.085.417 (per user) **Mean**: 69. 988, StdDev: 174.690
14	Regions travelled	Number of different regions in which photos have been taken, based on images' geoinformation.	**Total**: 532.426 (in the dataset) **Min**: 1, **Max**: 154 (per user) **Mean**: 12,09 **StdDev**: 13,18
15	Cameras Used	Different cameras used in photographing, according to images' Exif information.	**Total**: 5272 (in the dataset) **Min**: 1, **Max**: 267 (per user) **Mean**: 7,11 **StdDev**: 12,16
16	Photos	Number of uploaded photos per user.	**Total**: 922.487 (in the dataset) **Min**: 4, **Max**: 29.920 (per user) **Mean**: 1244,92 **StdDev**: 2714.2

4.2 Motivation and Community Feedback

Among other things, people are looking for approval and recognition. Positive feedback in one's photo uploads should be positively correlated to the rate of his/her uploads, as he/she would feel appreciated and valued within the community and would be inspired in sharing more content. In our study, we treat all comments as positive ones, since the vast majority of comments in Flickr are either to congratulate the uploader for his photo, or to propose him to participate in some private group.

We calculated the Pearson's correlation coefficient [25] for every user of the dataset. In Equation 2, X and Y stand for a user's total uploaded photos and comments received, x_i represents the number of user's uploaded photos at week i and y_i the comments received during this week, $\bar{x}$ and $\bar{y}$ are the mean values of X and Y and s_x and s_y are the standard deviations of X and Y.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad (2)$$

The mean value of Pearson's correlation coefficient for all users was found to be equal to 0.452. We hypothecate that there is some lag between the feedback received via comments and the effect they have in the alteration of the rate of photos uploads. To confirm this hypothesis we *"*slided*"* the comments data series, in an attempt to find when there is the maximum correlation with the photos uploads data series.

We defined a window of length w and, thus, we created w shifted versions of the initial comments series within the range of $-w/2$ and $w/2$. For $j \in [-w/2, w/2]$, we defined the shifted versions y originating from the initial series x as depicted in Equation (3).

$$z(i,j) = \begin{cases} y(i+j) & i+j > 0 \; AND \; i+j \leq n \\ 0 & i+j \leq 0 \; OR \; i+j > n \end{cases} \quad (3)$$

Having created the w shifted versions of the initial comments series, we searched for the specific j where the comments data series $Z(j)$ and the photos uploads data series X have the maximum correlation, as defined in Equation 4.

$$\begin{aligned} & maxCorrelation = \\ & = max\left(r_{xz(j)}\right) \\ & = max\left(\frac{\sum_{i=1}^{n}(x_i - \bar{x})\left(z_i(j) - \overline{z(j)}\right)}{(n-1)s_x s_z(j)}\right) \end{aligned} \quad (4)$$

After enough experimentation, we set up the length of the window to 60 ($|w| = 60$) and found $\bar{j} = 0.677$ for which the mean correlation coefficient value increased to 0.553. Thus, we could claim that the maximum effect of the comments received, appears 0.677 weeks after they are made.

The result presented in this section, indicate that since users have a small reaction time in positive feedback, the creation of a motivation mechanism for rewarding valued users could straightforwardly benefit WSM.

4.3 Effect of Event occurrence

It is common knowledge that events **do** play a crucial role in web user activities and that different people react independently in various events. Nevertheless, despite the advantages of understanding the effects of different world events to various groups of people, like the possibility to perform profile-based advertisement during the periods these events are on, the exact parameters of this action-result triggering model are not known, nor is its exact working mechanisms.

4.3.1 World Events

Within the context of our work we attempted to establish how popular world sports events affect people in general, and especially sports fans. For this reason we recorded a series of the most popular world events in the period 2003-2009, like the Olympics Games, the FIFA World Cups and the IAAF World Championships, as presented in Table 2.

Moreover, during the preprocessing phase we classified images as regular images and sports-related images. To do so, we first classified photo tags by selecting the tags that frequently appeared with the tag '*sport*'. For the set T of n tags, where the tag '*sport*' is represented as s, $s \subseteq T$, and the predefined threshold value v_1, we searched for the tags t_i, where $t_i \in T, |t_i \cap s| > v_1$. Having defined the sports-relevant tags, we classified as sports-related the ones tagged with many sports tags (for the set of P photos: $p_j \in P, |p_j \cap t_i| > v_2$). In our dataset, for $v_1 = 10$ and $v_2 = 2$, we identified 8962 photos relevant to sports, out of 922.487 photos.

Table 2. Summary of major sports event that took place from 2003 to 2009

Event	Country	Duration
IAAF Championship'03	France	23-31/8/2003
Summer Olympics 2004	Greece	13-29/8/2004
IAAF Championship'05	Finland	6-14/8/2005
Winter Olympics 2006	Italy	10-26/2/2006
FIFA World Cup 2006	Germany	9/6-9/7/2008
IAAF Championship'07	Japan	24/8-2/9/2007
Special Olympics 2007	Chine	2-11/10/2007
Summer Olympics 2008	Chine	8-24/8/2008
IAAF Championship'09	Germany	15-23/8/2009

Figure 4 depicts the fluctuation of the frequency of these sports photos with respect to all photos gathered. In the same figure, Table 2 sports events have been marked. One can easily distinguish that in times when popular world sports events take place the upload of photos related to sports increases.

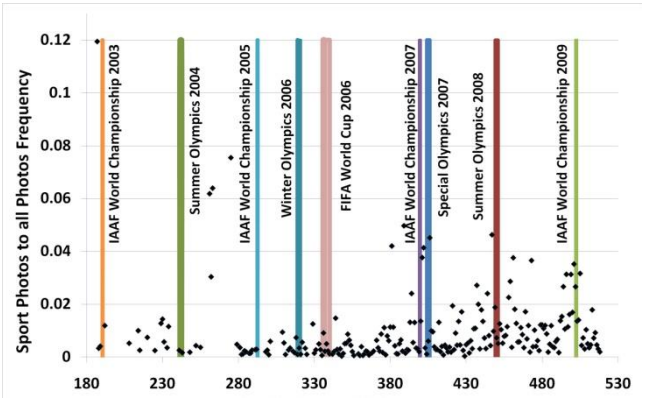

Figure 4: Ratio of photos relative to sports to all photos per week (July 2003 - December 2009)

In order to quantify the increase, we calculated the average upload of sports photos in regular periods (27.42 photo uploads/week) and during or near an event (44.89 photo uploads/week), an increase of 63%. *Near an event* means that we extended the duration of each event by 3 weeks, due to the reason that many users take some time to upload the photos they have taken (as discussed in Section 4.2). We have also taken into account photos posted after week 170, due to the limited use of Flickr before 2003 (our sample data were limited).

4.3.2 Personal Events

Personal events affect user activity as well. Although it is difficult, if not impossible, to know when a personal event like a birthday party, a trip or a meeting have occurred or exactly how these event have affected one's online activity, we hypothecate that we can model these events by examining the date photos were taken, as it is usual a peak to appear in the number of photos taken during these events.

The same type of analysis as in Section 4.2 was performed for finding out the correlation between the date photos are taken and the date they are uploaded in Flickr. The respective data series have already been presented in Figure 3. The Pearson's correlation coefficient was calculated to be 0.469 and for a length of the window w equal to 60 we found the mean movement of the photo taken data series equal to $\bar{J} = 3.17$ for which the correlation coefficient increased from 0.469 to 0.608, thus we could claim that users on average upload their photos 3 weeks after they have taken them.

4.4 Classification Models

Users' value within a content sharing community is directly linked to the quantity of content they share, and in the Flickr case, the number of photos they upload. Apart from the quantity of photos users share, the sharing rate is also important, as this can usually infer differences in personality and motivational characteristics [19].

Our goal was to use data mining techniques [23] in order to successfully predict each user's class, first with respect to the number of uploaded photos and, then with respect to the proportion on his/her active participation time by the total time being a member of the online community.

In order to predict the number of uploaded images, we trained a number of different classification models. We experimented with various families of algorithms, like trees, Bayesian algorithms, functional algorithms, lazy algorithms and classification models based on rule induction. A summary of the estimation accuracy achieved in terms of true positive (TP) classification rates and F-Measure with each classification algorithm is provided in Table 3, after performing 10-fold cross validation.

As Table 3 depicts, the logistic regression model using ridge estimators [15] achieved 72.74 true positives accuracy and was the best performing model. The detailed results for this model are available in Table 4.

Table 3: Comparison of Different Algorithms for Classifying Users According to Uploaded Photos

Family	Algorithm	Accuracy	F-Score
Function	Logistic Regression	72.74%	0.725
Function	Platt's sequential optimization algorithm	71.39%	0.712
Trees	Alternating LogitBoost	69.37%	0.677
Rules	Decision Table	69.10%	0.686
Trees	Best-first Decision Tree	67.61%	0.677
Function	Normalized Gaussian Radial Basis Network	67.20%	0.666
Bayes	Naïve Bayes	66.94%	0.665
Bayes	NB: Aggregating One-Dependence Estimators	69.64%	0.692
Rules	PART Decision List	67.34%	0.670
Trees	Forest of Random Trees	66.94%	0.666
Rules	Repeated Pruning for Error Reduction	66.94%	0.630
Trees	C4.5 Tree	65.86%	0.648
Rules	Ripple-Down Rule	64.51%	0.665
Lazy	K-nearest neighbors	64.10%	0.634
Lazy	K* Instance Based	62.75%	0.626

The model reached 72.74% accuracy outperforming all the other models tested. May one take a closer look at the confusion matrix and considering the ordinal nature of the data, one can assume that the results are fairly close to reality. After all, the discrimination of users to '*Low*', '*Medium*' and '*High*' uploaders is a bit cumbersome; in real life it is hard to define a threshold between the '*Low*' and '*Medium*' and between the '*Medium*' and '*High*' uploader. This discrimination is even harder considering the power law distribution of the number of photos users have uploaded. The power law distribution is better exhibited by the data presented in Table 1: the 741 users of the dataset have uploaded 922.487 photos and each user has uploaded from 4 (min) to 29.920 (max) photos. The mean value of photo uploads per user if 1244, while the standard deviation is 2714.

Table 4: Total Number of Uploaded Photos

Class	**Classified As**		
	Low	**Medium**	**High**
Low	207 (27.94%)	40 (5.40%)	1 (0.14%)
Medium	49 (6.61%)	139 (18.76%)	59 (7.96%)
High	2 (0.27%)	51 (6.88%)	193 (26.05%)
Validation Results			
True positives:	**539 (72.74%)**		
Precision:	0.723		
Recall:	0.727		
F-Measure:	0.725		

In a similar manner, we trained a model predicting users' ratio of active time to the total time since their first photo upload, as computed by Equation (1).

$$weeks\ uploading\ photos/weeks\ since\ first\ upload \quad (1)$$

Table 5 summarizes the accuracy result achieved from a number of different models in classifying users according to the ratio of active to total participation time.

Table 5: Comparison of Different Algorithms for Classifying Users According to Ratio of Active/Total Participating Time

Family	Algorithm	Accuracy	F-Score
Function	Platt's sequential optimization algorithm	59.78%	0.593
Function	Logistic Regression	58.43%	0.580
Rules	Ripple-Down Rule	54.93%	0.536
Trees	Alternating LogitBoost	53.58%	0.532
Function	Normalized Gaussian Radial Basis Network	53.44%	0.532
Bayes	NB: Aggregating One-Dependence Estimators	54.39%	0.538
Trees	C4.5 Tree	52.23%	0.518
Trees	Best-first Decision Tree	51.15%	0.500
Bayes	Naïve Bayes	51.15%	0.499
Rules	PART Decision List	51.01%	0.510
Rules	Decision Table	50.47%	0.495
Trees	Forest of Random Trees	50.33%	0.501
Rules	Repeated Pruning for Error Reduction	50.06%	0.497
Lazy	K-nearest neighbors	47.10%	0.470
Lazy	K* Instance Based	45.34%	0.454

The model that achieved the best accuracy was Platt's sequential minimal optimization (SMO) algorithm with the polynomial kernel: $K(x,y) = (< x,y > +1)^{\wedge}p$, [13],[24]. As depicted in Table 6, Platt's sequential minimal optimization (SMO) algorithm reached 59.78% accuracy outperforming all the other models tested. May one take a closer look at the confusion matrix and considering the ordinal nature of the data as well as their power law distribution (Mean=0.078, Standard Deviation=0.056, Minimum=0, Maximum=1), one can assume that the results are, again, fairly close to reality.

Table 6: Periods uploading photos to the total period of being a member

Class	Classified As		
	Low	**Medium**	**High**
Low	181 (24.43%)	54 (7.29%)	12 (1.62%)
Medium	65 (8.77%)	105 (14.17%)	77 (10.39%)
High	33 (4.45%)	57 (7.69%)	157 (21.19%)
Validation Results			
True positives:	**443 (59.78%)**		
Precision:	0.591		
Recall:	0.598		
F-Measure:	0.593		

4.5 Association Rules

Finally, we performed association analysis, hoping to discover interesting association rules and patterns. We used an Apriori-based algorithm [1], [11]. Although a lot of association rules can be extracted (with respect to the minimum confidence and support selected) most of them being trivial. For example, it is expected that users without a Pro account would have uploaded a small number of photos, since there is a limit of 200 images, or that users with a lot of uploaded photos have also a lot of geotagged photos.

Nevertheless, some interesting rules were discovered (presented in

Table 7). For example, according to Rule-5, even though some users have uploaded a "medium" number of photos, if they have a large number of contacts they appear to have a high number of comments per photo. We remind that the contact relationship in Flickr is not reciprocal, so users with a lot of contacts are not necessary contacts of a lot of people. Rule-8 is also interesting: even if some people are not old members of Flickr, they ratio of comments per photo may be high in case they have declared a lot of favorite photos. Of course, all these rules have to be closely examined and tested on a larger dataset in order to confirm their validity, but nonetheless they comprise an inspiring start for further analysis.

5. CONCLUSIONS AND FUTURE WORK

In this work we have presented a conceptual model towards understanding how personality, motivations and events affect web activity in Social Media. We focused on content creation and sharing on Flickr, an online community for sharing images.

We showed that events greatly affect particular types of photo uploads. In the analysis of sports events we estimated an increase of 63% in the rate of sports photos uploads during the periods of major world sports events. We attempted to predict the rate and volume users share content on the web based on publicly online available information. The accuracy of our model regarding the classification of users with respect to their uploaded images was 72.74%, while in respect to their active time in the website was 59.78%. Taking also into account the ordinal nature of data and looking carefully in the confusion matrix we claim that these success rates are even greater.

We calculated the correlation between comments received from the community and the rate users upload photos. The correlation is maximized if we move the comments data series by 5 days, thus we claim that comments received have the maximum effect some (5) days. Likewise, we calculated the correlation between the dates photos are taken and the dates they are posted on Flickr. We found that this correlation is maximized if we slide the date taken data series by 3 weeks, thus we claim that on average users upload their photos after a period of some (three) weeks.

We also performed association analysis on the profile elements and behavior of Flickr users. Though a lot of the association rules derived are trivial, others present special interest and are inspiring for future analysis.

Future work includes more extensive validation of the results in bigger datasets, for larger time intervals. Analysis of other content sharing websites like YouTube, Panoramio and LiveJournal is also programmed. Analysis and prediction of the fluctuation of the number of photos users share each week will be performed using time series and signal processing techniques.

Moreover, our future plans include the study on the effects of different kinds of world events compared to different profiles regarding content creation and sharing as we would like to augment the sports event of this study with politic, environmental, medical, scientific and other events. Finally, the further study of

Table 7: Summary of interesting association rules

Variable 1	Variable 2		Result	Support	Confidence
Contacts = High	-	→	Favorites = High	234 (31.58%)	0.79
Comments/Photo = High	-	→	Pro User	247 (33.33%)	0.79
Comments/Photo = Low	-	→	Groups = Low	247 (33.33%)	0.73
Contacts = High	-	→	Comments/Photo = High	234 (31.58%)	0.72
Uploaded Photos = Medium	Contacts = High	→	Comments/Photo = High	78 (10.53%)	0.84
Uploaded Photos = Medium	Favorites = High	→	Comments/Photo = High	88 (11.88%)	0.84
Uploaded Photos = Medium	Contacts = High	→	Favorites = High	77 (10.39%)	0.83
Participating Weeks = Medium	Favorites = High	→	Comments/Photo = High	85 (11.47%)	0.85

association rules and the exact determination of their statistical significance are expected in our future work.

As we are shifting towards Web 3.0, electronic commerce is transforming from trading materials and services into trading content. In the core of new promising technologies of the Web, like the Semantic Web, Linked Data, and Social Media are lying on content interconnection and cognitive content. It is clear that users are now responsible not only for content consumption, but for content creation as well. Consequently, it is vital to motivate users to create and share content in the Web and in order to so, one has to understand why, when and how online content is created and interacted upon.

6. REFERENCES

[1] Agrawal, R. and Srikant R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In Proceeding of the 20th International Conference on Very Large Data Bases (September 12-15, 1994, Santiago de Chile, Chile), 478-499.

[2] Allport, G. W. 1937. Personality: A psychological interpretation, Henry Holt.

[3] Allport, G. W. 1963. Pattern and Growth in Personality, Harcourt College Publishers, ISBN: 0030108101.

[4] Allport, G. W. 1985. The historical background of social psychology. In Lindzey G. and Aronson E. (Eds.), The handbook of social psychology. New York: McGraw Hill, ISBN: 0195213769.

[5] Bakshy, E., Karrer, B., and Adamic, L. A. 2009. Social influence and the diffusion of user-created content. In Proceedings of the Tenth ACM Conference on Electronic Commerce (Stanford, California, USA, July 06 - 10, 2009). EC '09. ACM, New York, NY, 325-334. DOI= http://doi.acm.org/10.1145/1566374.1566421

[6] Bhat, S., Bevans, M., and Sengupta, S., 2002. Measuring Users' Web Activity to Evaluate and Enhance Advertising Effectiveness, *Journal of Advertising*, 31(3): 97-106

[7] Cha, M., Kwak, H., Rodriguez, P., Ahn, Y., and Moon, S. 2007. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In Proceedings of the 7th ACM SIGCOMM Conference on internet Measurement (San Diego, California, USA, October 24 - 26, 2007). IMC '07. ACM, New York, NY, 1-14. DOI= http://doi.acm.org/10.1145/1298306.1298309

[8] Flickr Blog, October 12, 2009, 4.000.000.000 http://blog.flickr.net/en/2009/10/12/4000000000/, Retrieved 2010-01-04

[9] Flickrj, http://sourceforge.net/projects/flickrj/, Retrieved 2010-01-04

[10] Freeman, L. C. 2004. The Development of Social Network Analysis: A Study in the Sociology of Science. Empirical Press. ISBN: 1594577145.

[11] Hu, K., Lu, Y., Zhou, L., and Shi, C. 1999. Integrating Classification and Association Rule Mining: A Concept Lattice Framework. In Proceedings of the 7th international Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing (November 09 - 11, 1999). N. Zhong, A. Skowron, and S. Ohsuga, Eds. Lecture Notes In Computer Science, vol. 1711. Springer-Verlag, London, 443-447.

[12] Huberman, B. A., Romero D. M., and Wu F., 2009. Crowsourcing, Attention and Productivity. Journal of Information Science 35(6): 758-765.

[13] Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. 2001. Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Comput. 13, 3 (Mar. 2001), 637-649. DOI= http://dx.doi.org/10.1162/089976601300014493.

[14] Lambel, J. and Bhalla, A., 2007. The role of status seeking in online communities: Giving the gift of experience, Journal of Computer-Mediated Commuinication 12(2) 435-455

[15] le Cessie, S., and van Houwelingen, J.C. 1992. Ridge Estimators in Logistic Regression. Applied Statistics 41(1): 191-201.

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

[17] McAdams, D. P. 2005. The Person: A New Introduction to Personality Psychology, Wiley, ISBN: 0471716995.

[18] Milgram, S. 1967. The small world problem. Psychology Today, 2:60–67, 1967.

[19] Miller, A. D., and Edwards, W. K. 2007. Give and take: a study of consumer photo-sharing culture and practice. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03, 2007). CHI '07. ACM, New York, NY, 347-356.

[20] Nov, O. 2007. What Motivates Wikipedians?, Communications of the ACM, 50(11): 60-64.

[21] Nov, O.; Naaman, M. and Ye, C. 2008, Motivational, Structural and Tenure Factors that impact Online Community Photo Sharing. International Conference on Weblogs and Social Media (ICWSM 2008), May 2009, San Jose, California.

[22] O'Reilly T. 2005. What Is Web 2.0., O'Reilly Network. http://oreilly.com/web2/archive/what-is-web-20.html , Retrieved 2010-01-04.

[23] Pang-Ning T., Steinbach M., and Kumar V. 2005. Introduction to Data Mining, Addison Wesley, ISBN: 0321321367.

[24] Platt, J., 1998, Machines using Sequential Minimal Optimization. In Schoelkopf, B., Burges, C., and Smola, A. editors, Advances in Kernel Methods - Support Vector Learning, MIT Press, ISBN: 0262194163.

[25] Rodgers, J. L., Nicewander, W. A. 1988. Thirteen ways to look at the correlation coefficient. The American Statistician, (Feb 1988),42(1):59–66

[26] Sellen, A. J., Murphy, R., and Shaw, K. L. 2002. How knowledge workers use the web. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Changing Our World, Changing Ourselves

(Minneapolis, Minnesota, USA, April 20 - 25, 2002). CHI '02. ACM, New York, NY, 227-234. DOI= http://doi.acm.org/10.1145/503376.503418

[27] Scott, J. P. 2000, Social Network Analysis, Sage Publications, ISBN: 0761963391.

[28] Technical Standardization Committee on AV & IT Storage Systems and Equipment, Exchangeable image file format for digital still cameras: Exif Version 2.2, http://www.digicamsoft.com/exif22/exif22/html/exif22_1.htm, Retrieved 2010-01-04

[29] Wasko, M., and Faraj, S., 2005. Why Should I Share? Examining Social Capital and Knowledge Contribution in Electronic Networks of Practice, MIS Quarterly, 29(1): 35-37.

[30] Yang, J., Adamic, L. A., and Ackerman, M. S. 2008. Crowdsourcing and knowledge sharing: strategic user behavior on taskcn. In Proceedings of the 9th ACM Conference on Electronic Commerce (Chicago, Il, USA, July 08 - 12, 2008). EC '08. ACM, New York, NY, 246-255. DOI= http://doi.acm.org/10.1145/1386790.1386829