Wearing a YouTube Hat: Directors, Comedians, Gurus, and User Aggregated Behavior

Joan-Isaac Biel
Idiap Research Institute
EPF Lausanne
Switzerland
jibiel@idiap.ch

Daniel Gatica-Perez
Idiap Research Institute
EPF Lausanne
Switzerland
gatica@idiap.ch

ABSTRACT

While existing studies on YouTube's massive user-generated video content have mostly focused on the analysis of *videos*, their characteristics, and network properties, little attention has been paid to the analysis of *users' long-term behavior* as it relates to the roles they self-define and (explicitly or not) play in the site. In this paper, we present a novel statistical analysis of aggregated user behavior in YouTube from the novel perspective of user categories, a feature that allows people to ascribe to popular roles and to potentially reach certain communities. Using a sample of 270,000 users, we found that a high level of interaction and participation is concentrated on a relatively small, yet significant, group of users, following recognizable patterns of personal and social involvement. Based on our analysis, we also show that by using simple behavioral features from user profiles, people can be automatically classified according to their category with accuracy rates of up to 73%.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services

General Terms

Measurement, Human Factors

1. INTRODUCTION

Social media sites have become mainstream publishing and communication tools that have globally changed media production and consumption patterns. Among them, YouTube is one of the best examples of this explosion of online user-generated content, receiving 20h of new videos every minute (the equivalent of 115,000 Hollywood movies per week) [1], and concentrating over 40% of all online video viewers and 60% of all watched videos [5]. While the first key achievement of YouTube was the creation of an easy-to-use integrated platform aimed to upload, share, and watch online videos, removing barriers to the online video scene, it further promoted participation allowing users to comment, rate, explore, and post related videos. As a result, YouTube has become a site for people and communities to join and interact, from aspiring rockstars to top politicians.

Recently, research on YouTube has analyzed the statistics and social network of videos, revealing properties of the nature of video sharing systems that may be key for the future of such services [3, 4, 8]. However, few works have focused on characterizing long-term user behavior. Halvey and Keane provided a brief statistical analysis on the use of YouTube social-oriented features [6], Santos et al. studied the properties of the social network of friends and subscriptions [7], and Benevenuto et al. presented a similar characterization based on user video interactions [2]. Understanding how users typically behave and interact, and how they perceive YouTube as both a system and a social outlet is fundamental in order to improve its performance. However, existing work has treated YouTube users as a single homogeneous group, ignoring potential differences between groups of users. We believe that this is an essential point when analyzing users in large social networks, due to the likely wide variety of behavioral patterns.

This paper, to the best of our knowledge, is one of the the first attempts to analyze large-scale aggregated behavior of YouTube users under the lens of *user categories*, i.e., self-assigned roles that people can choose among *Director*, *Comedian*, *Guru*, *Musician*, or *Reporter*. We hypothesize that users' choices and the implicit or explicit ways in which people respond to these labels give rise to different collective behavior that can be measured quantitatively. Our paper has two contributions. First, on a large user dataset, we present a statistical analysis of YouTube users and categories based on easy-to-extract, long-term behavioral features from users, which do not require any video or metadata processing. Our analysis reveals clear trends regarding people's category choices, and differences on user behavior (both individual and social) across categories and gender, which suggest that the emerging communities do have differences that could lead the way to automatic modeling of groups of users. Second, we use such behavioral cues in various classification tasks, in order to explore whether, together or alone, they can be used to predict user categories, obtaining promising performance. Overall, our work aims at complementing the emerging (and much needed) work in sociology and ethnography on the understanding of users' motivations to select roles and to create and maintain self and group identities in social media outlets like YouTube, and enquires about some fundamental needs for personalized applications.

MM'09, October 19–24, 2009, Beijing, China.

2. USER CATEGORIES IN YOUTUBE

The YouTube platform is in continuous transformation. As its community grows, diversifies, and creates new trends of interaction, new features are added to the site. As such, special user categories have been gradually introduced, originally serving different purposes.

In April 2006, YouTube introduced the *Director* program in response to a video length limitation earlier installed to prevent copyright infringement, more likely brought about by long videos. A user proving to be a legitimate creator of his/her uploaded content could apply for a *Director* account that allowed to upload videos longer than 10min[1]. In a span of five months, YouTube created special accounts for other users willing to promote their work. Users with a *Musician* or a *Comedian* accounts had the possibility to customize their user profiles publishing performer information and a schedule of show dates. By June 2008, four more accounts had been introduced. *Guru* and *Reporter* accounts were respectively addressed to people devoted to create "how to" videos (i.e. videos that teach certain skills or explain how to do something) and to people dedicated to inform others about news and events occurring around them. While these accounts allowed anyone to sign up, *Non-profit* and *Politician* accounts were only to be held by non-profit organizations and politicians. The first one aimed to support advocacy campaigns and fundrising efforts. The second one was created for candidates of the 2008 United States presidential election and some other elections.

Today, a new user signing up for a YouTube account receives by default a *Standard* or *YouTuber* account, with the basic YouTube features such as uploading, commenting, etc. (we use the term *Standard* to avoid confusing *YouTubers* with all the YouTube users). This status remains unchanged unless the user *intentionally* modifies his channel type to one of the special categories[2]. In doing so, he is allowed a certain level of customization on his channel, where he also exhibits a label with the name of the user category.

Why would a user be interested in becoming a special user? And why would he choose one category or another? A potential benefit for special users is that they qualify for the "Channels" page in YouTube, where the most subscribed and the most viewed special user channels are featured. Compared to browsing and searching, video and channel promotion in YouTube are advantageous ways of attracting viewers. However, one may argue that this only benefits a small set of users, which then accumulate a high number of subscriptions and views. Moreover, though the different user categories are described in the YouTube help section, users are free to assign a user category to themselves independently of how well they "fit" the respective category description, possibly augmenting the overlap among categories. For some users, like *Musicians*, the answer may be on the goodness of fit under a specific description. For others, we hypothesize that the process of self-assigning a specific user category conveys a sense of belonging to a specific community of users. Thus, a user would tend to acquire the same user category than the users he follows or interacts with, somehow reassuring his presence in that community.

[1]Eventually, new *Directors* lost such privilege.

[2]A user *channel* is the YouTube equivalent to a user *profile*. In addition to user related information, it contains the videos uploaded.

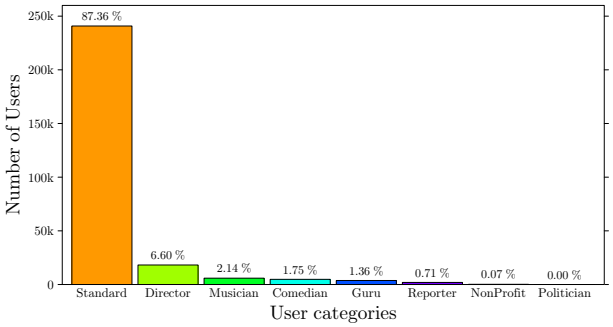

Figure 1: Distribution of user categories

3. ANALYSIS OF USER CATEGORIES

Using YouTube's API, we followed a two-step data collection procedure that consisted on first obtaining the last uploaded videos from the site, extracting the username of the uploader, and then retrieving the channel information from every user. We repeated this procedure every 5min, from March 5th to March 9th, 2009. We used video-category specific queries to overcome a 999 entries per feed limitation existing in the API feeds, thus augmenting our capacity to obtain the last uploaded videos.

We collected a dataset of 273,000 distinct users, for whom we obtained a set of descriptive behavioral features. We explicitly limited the set of features to all the attributes that can be easily extracted from the users' profile, aware that there might be other features that are richer descriptors of behavior at higher computational cost. We divided the behavioral features in two different groups, capturing the individual participation and the social-oriented behavior of users, respectively.

The **individual participation features** are direct indicators of individual user activity in the site, namely **uploads**, **videos watched**, and **favorites**. The first one is the number of videos uploaded by the user, which is a measure of his contribution to the website content. The second one is a measure of the participation of the user as a viewer. This might be a noisy measure, since it does not take into account the videos watched when the user is not logged in. The last feature is the number of videos marked as favorite, which indicates that the user viewed a video and liked it. This may be a descriptor of the way the user further engages with videos.

The **social oriented features** are measures of an interaction between users. We differentiate between incoming or outgoing features, depending on the role of the users in the interaction. Incoming features describe how other users perceive the user, and include the accumulated number of **views**, and the number of **subscribers**. The first one is a measure of popularity in the YouTube community. Note, however, that the number of views could be biased to few highly viewed videos. The second one is related to the interest of other users in the videos created by the user. A subscribed user receives a notification whenever the user he subscribed to has uploaded new material. Outgoing features reveal the interest of the user towards other users, and consist on the number of **subscriptions** and the number of **friends**. The first one is an indicator of the interest on the videos of other users. The second one denotes an interest on other users which may go beyond the videos uploaded, and might refer to "true" friendships.

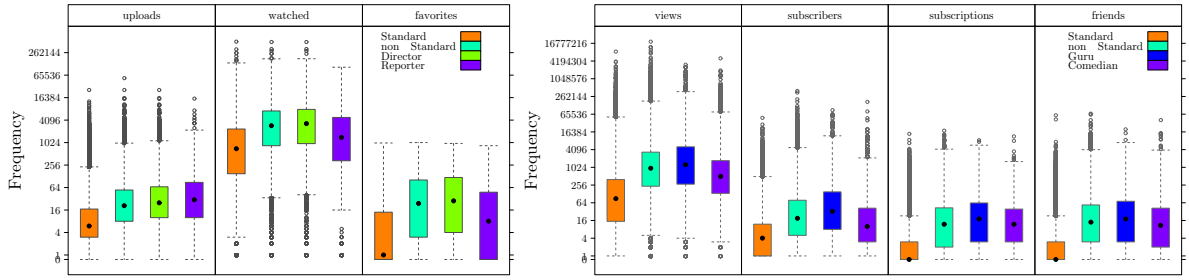

Figure 2: Boxplot of participative features (left) and social-oriented features (right) for several user categories.

3.1 Are you special?

First of all, we are interested in exploring the level of popularity that user categories have among the YouTube community. As shown in Figure 1, our data reveals that 12.6% of the users are choosing to label themselves with a category different than *Standard*. The relatively moderate level of popularity of special categories could be explained by the poor advertising of such feature in the site. We hypothesize that users are more likely to find about categories by "word of mouth", after seeing other users with labels such as *Director* or *Comedian*.

The distribution of special categories seems to be biased by the chronological order in which categories were introduced. This might suggest, in fact, that the process of choosing a category is influenced by a "rich-get-richer" effect, in which new users would tend to choose the most numerous category. This means that *Director*, the most popular among the special categories, is probably bringing together a larger variety of different users and behaviors, compared to *Comedians*, *Musicians* or *Gurus*.

3.2 How participative are you?

As it was exposed in earlier work [6], the number of uploads, videos watched, and favorites per user indicate a relatively low level of individual participation in YouTube (see Figure 2, left). However, our category-based analysis reveals that this result is biased by the very low participation of *Standard* users, compared to special users. In median values, *Standards* uploaded 6 videos, watched 734 videos, and had "favorited" only 1 video, versus the 21 uploads, 2803 videos watched and 22 favorites of the special users. We argue that due to their level of engagement with the site, *Standards* are more likely than other users to view videos without being logged in, which might explain, at least partially, the big difference in the number of videos watched.

Among special users, *Directors* and *Gurus* are the most participative in all the aspects, followed by *Comedians* and *Musicians*. Instead, *Non-profit*, *Politicians*, and *Reporters* follow a different pattern of participation. Whereas they are also very active uploaders, they have a very low number of videos watched and "favorited", which may indicate that they are more interested in releasing their work or spreading their messages, rather than exploring the site's content.

3.3 How social are you?

Social-oriented features follow a similar pattern than participative features regarding the differences between *Standard* users and special users. As shown in Figure 2 (right), *Standards* accumulated, in median values, 89 views and 4 subscribers, versus the 958 views and 19 subscribers of special users. *Standards* have also no subscriptions or friends in median value. This likely has to do with the motivation behind YouTube users. *Standard* users, in general, might be more interested in sharing few casual videos with their relatives or friends, which would potentially generate a small number of views. Instead, special users seem to be more interested in interacting with the YouTube community through videos that are of wider interest among other users, thus receiving more views and subscriptions. We find different patterns of interaction. *Directors*, *Gurus*, *Non-profit*, and *Reporters* are the ones who get more attention from the community through both views and subscriptions. Compared to them, *Politicians* and *Comedians* accumulate similar number of views, but few subscribers. Surprisingly, *Musicians* are, among special users, the ones that receive less views and subscriptions. This differs from other social sites such as MySpace, where people promoting their music represent a highly active community.

Directors, *Comedians* and *Gurus* are also the most prominent user categories in terms of friends and subscriptions, which suggests that YouTube-specific contacts and "real life" contacts have a similar presence in their online interaction. The rest of the users have, generally, higher number of friends than subscriptions.

3.4 Male or female?

Gender analysis, and in particular, gender distribution, uncovers interesting aspects of the YouTube behavior. Based on our data, YouTube concentrates a higher participation number of men (73% of the users) than women (27%). Moreover, we find that male users are more likely to enroll in special user categories (13% vs. 9%).

Gender differences in the distribution of special categories are very small. However, the distribution of both participative and social-oriented features show very different patterns of behavior between men and women. Though men and women participate with a similar rate of uploads and videos watched per user, special female users "favorite" 82% more videos than men. This suggests a different pattern on the way women watch and engage with video content. In addition, women accumulate 22 % more subscribers, have 75% more subscriptions, and double the median number of friends that men have, which suggests that women, overall, have a more social-driven behavior in YouTube than men.

Some of these findings are not completely new but are backed up by substantially more data. In a sociological manually-driven analysis of a very small random sample of 100 YouTube vlogs, Molyneaux et al. [9] found a higher presence of male users. They also found that women were most likely to interact with the YouTube community through their videos, and that they receive a higher number of views than men.

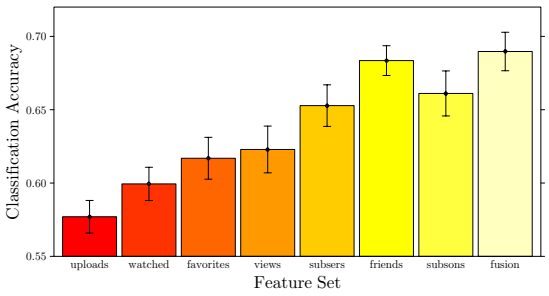

Figure 3: *Standard* vs. non-*Standard* CAR using behavioral features alone and together.

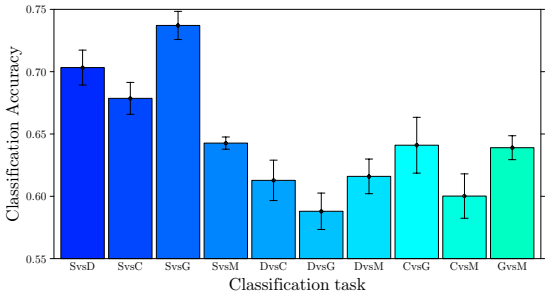

Figure 4: Best CARs on binary tasks (capital letters correspond to the initials of the categories' names).

4. CLASSIFYING YOUTUBE USERS

The analysis of the previous section suggests that the basic features in the YouTube users' profile could be used to characterize user categories. In order to explore the goodness of this characterization, we defined a series of classification tasks where a YouTube user is classified between two given user categories using different combinations of features. We use the results not only to evaluate the discriminative power of the features, but to measure the similarity between the behavior of different types of users.

Each task consisted on a 10-fold cross-validation using a SVM classifier with a Gaussian Kernel. In training, we optimized σ and C using 5-fold cross-validation.

Our first binary classification task was between *Standard* and non-*Standard* users (as one unique category) on a balanced subset of 10,000 users randomly selected from our dataset. We only considered the most popular special categories: *Directors*, *Musicians*, *Comedians* and *Gurus*, which were also balanced among the 5,000 corresponding samples. Results in Figure 3 show an average classification accuracy rate (CAR) of 68.9% $\pm$ 1.2 for the fusion of features (random performance is 50%). However, we found that using the number of friends as a single feature one could achieve a classification rate of 68.3% $\pm$ 1.4. This might indicate that differences in user behavior between *YouTubers* and special users, are to be found mainly in the level and pattern of interaction with other users, that is, in how social they are. In fact, that would also explain why subscriptions and subscribers are the next best single features.

The rest of the binary tasks were defined between pairs of single user categories (see Figure 4), using 5,000 users per category. Data used on tasks involving categories with less than 5,000 users (e.g. *Gurus*) were balanced to the number of users of such categories. Except for those tasks involving *Standards*, the best CARs were obtained with several combinations (of at least three) social-oriented features, indicating higher similarity among special categories. Single features would drop the best CAR to 53%. For tasks involving *Directors* and *Gurus*, the number of views appeared in almost all winning combinations, whereas for *Comedians* and *Musicians* different combinations led to similar results. In some cases, replacing only one of the social features for a participative feature would not cause a drop in performance, suggesting that the latter features still contain different patterns of behavior between special categories.

The features obtained from the users' channel are capturing broad statistics on different aspects of the long-term users' behavior. Finer aspects could be obtained by extracting other features in a more computationally expensive manner, which could hopefully lead to a better characterization. As an example, new participative features could include the frequency of the uploads, the number of comments posted, or user-specific videos' features, including metadata and audio-visual features. Social-oriented features such as friends, subscriptions, or subscribers could be further divided in inter-category and intra-category relations. In addition, considering the antiquity of YouTube users could help to erase some dissimilarities between users of the same behavioral group.

5. CONCLUSIONS

In this paper, we have presented an analysis of YouTube users' long-term behavior using easy-to-extract features from the users' channels and large-scale data. We have shown that the group of special YouTube users is an actively social community, as opposed to *Standard* users, and we have revealed different patterns of participation and interaction and highlighted the possible motivations behind them. Furthermore, we have backed up earlier findings on gender behavioral division using (by 3 orders of magnitude) more data. A series of binary classification tasks between YouTube users categories has shown social-oriented features to be key when describing differences between user groups' behaviors. The use of more computationally expensive features to capture finer aspects of the users' behavior will be the subject of our future work.

Acknowledgments: we thank the financial support provided by the Swiss National Center of Competence (NCCR) on Interactive Multimodal Information Management (IM)2.

6. REFERENCES

[1] YouTube Blog, May 20, 2009.
[2] F. Benevenuto, F. Duarte, T. Rodrigues, V. Almeida, J. M. Almeida, and K. W. Ross. Understanding video interactions in YouTube. In *Proc. of 16th ACM MM*, Oct. 2008.
[3] 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 *Proc. of the 7th IMC*, Oct. 2007.
[4] X. Cheng, C. Dale, and J. Liu. Statistics and social network of YouTube videos. In *Proc. of the 16th IWQoS*, May 2008.
[5] Comscore. Press Realease, Dec. 9, 2008.
[6] M. Halvey and M. Keane. Exploring social dynamics in online media sharing. In *Proc. of the 16th WWW*, May 2007.
[7] R. L. Santos, B. P. Rocha, R. G. Rezende, and A. A. Loureiro. Characterizing the YouTube video-sharing community. Technical report, 2007.
[8] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *Proc. of the 7th ACM SIGCOMM IMC*, 2007.
[9] H. Molyneaux, S. O'Donnell, K. Gibson, and J. Singer. Exploring the gender divide on YouTube: an analysis of the creation and reception of vlogs. *AC Journal*, 10(2), 2008.