Identifying User Behavior in Online Social Networks*

Marcelo Maia, Jussara Almeida, Virgílio Almeida
Computer Science Department
Federal University of Minas Gerais
Av. Antônio Carlos, 6627, Pampulha
Belo Horizonte, MG, Brazil, 31270-010
{mmaia, jussara, virgilio}@dcc.ufmg.br

ABSTRACT

Online social networks pose an interesting problem: how to best characterize the different classes of user behavior. Traditionally, user behavior characterization methods, based on user individual features, are not appropriate for online networking sites. In these environments, users interact with the site and with other users through a series of multiple interfaces that let them to upload and view content, choose friends, rank favorite content, subscribe to users and do many other interactions. Different interaction patterns can be observed for different groups of users. In this paper, we propose a methodology for characterizing and identifying user behaviors in online social networks. First, we crawled data from YouTube and used a clustering algorithm to group users that share similar behavioral pattern. Next, we have shown that attributes that stem from the user social interactions, in contrast to attributes relative to each individual user, are good discriminators and allow the identification of relevant user behaviors. Finally, we present and discuss experimental results of the use of proposed methodology. A set of useful profiles, derived from the analysis of the YouTube sample is presented. The identification of different classes of user behavior has the potential to improve, for instance, recommendation systems for advertisements in online social networks.

Categories and Subject Descriptors

C.2.3 [**Network Operations**]: Network monitoring

General Terms

Human Factors, Measurement, Experimentation, Theory

Keywords

Online Social Networks, Clustering, Groups, User Behavior

*This research was sponsored by UOL (www.uol.com.br), through its UOL Bolsa Pesquisa program, process number 20060520221328a.

SocialNets'08, April 1, 2008 , Glasgow, Scotland, UK

1. INTRODUCTION

Enabling and encouraging interactions between users have shown to be of paramount importance to the triumph of most of the online services available on the current Internet. Often referred to as Web 2.0, some of the most successful services like Flickr[1], MySpace[1], Orkut[1] and YouTube[1] allow users to create and maintain their own personal web logs (or *blogs* for short), which can be comprised of text, pictures and videos. Users of these services typically execute different kinds of tasks such as search for information, watch or view public content, exchange messages, among others.

Due to the intrinsic human nature, users are not expected to exhibit one single and simple behavior. Some are enthusiasts and express themselves by updating blogs on a daily basis and upload as many videos or photos as they can, whereas there are users that act like free-riders [1, 10] and just want to enjoy the contents made publicly available.

The comprehension of peculiarities of the user behavioral patterns in online social communication systems are of significant importance in the sense they can give useful information for the providers to plan their resource capacity such as network bandwidth usage, memory allocation and CPU. To aid specialists in their designing tasks, user behavior models have also been extensively studied and are able to capture particularities from a community of users with a single behavior to multiple classes of users [2, 3, 6, 11, 18].

In order to develop models that capture accurately the different user behavior classes, specialists must know how to identify and distinguish users with different aspirations and purposes. This work seeks to systematically identify user behaviors in online social systems. Overall, to achieve this goal we designed a web crawler to gather data from YouTube subscription network, where users are linked if they subscribe to videos of other users. The crawled data were used to build one *feature vector* per user composed of a set of individual and social information relative to the user. The feature vectors were given as input to the K-means clustering algorithm [9, 13, 15], which in turn grouped users with similar behavior. We provide the details in the following sections.

The main contributions of this work are:

- The analysis of YouTube's subscription network, which is more suitable to capture the process of production, promotion and visualization of user generated content.
- We have shown that the individual attributes of the users are not good behavior discriminators because

[1]http://www.{flickr,myspace,orkut,youtube}.com

they result in one single large group. On the other hand, the attributes stemmed from the social interactions are those that allow the distinction of the groups.

- The methodology used in this paper for the user behavior identification can be applied not only to YouTube but also to online social networking systems in general.

The remaining of this paper is organized as follows. Section 2 presents the related work. The data collection is shown in section 3. The methodology is described in section 4. Section 5 presents the behavior identification process and section 6 concludes this work and exposes future directions.

2. RELATED WORK

A number of studies focusing on properties of the Web 2.0 have been recently published. They analyze aspects of the users, the structure of the social networks emerged from their interactions and also show how this knowledge can help to design new systems or improve existing ones.

The authors in [8] have studied properties of the user generated videos such as popularity shifts. Video traffic and file attributes, like bitrate, were analyzed in [12]. Structural properties of the friendship network of three major systems were compared in [4]. The authors in [19] presented a large-scale measurement study and analysis of the structure of four major systems in which they have confirmed the power-law, small-world and scale-free properties of the services.

Tapping into the user behavior, the authors in [20] presents a characterization of broadband user behavior from an Internet Service Provider standpoint. The authors in [6] analyzed a large online community system and investigated engagement and relationship between users and groups of users. In [11] it is presented a method for applying social network analysis to characterize authors in a newsgroup system. In [18] the authors considered classes of users to optimize the computational resources and the e-commerce business itself. In [2] the user behavior was incorporated to significantly improve the accuracy of a web search ranking algorithm. In [3] the authors evaluated a general classification framework they have designed based on social feedback to separate high quality question-answer items from the rest.

Our work differs fundamentally from the aforementioned references. Rather than analyzing the user generated content or the structure of the networks emerged from their interactions, we investigate typical user behaviors and present a methodology to identify characteristics that define a user as part of a group that share the same overall behavior.

3. DATA COLLECTION

Founded in 2005, YouTube is the largest online social system for sharing video clips on the current Web 2.0 [5, 7]. Due to its magnitude it has been chosen for the profile analysis. Videos within YouTube are assigned to 12 different categories and can be browsed by 9 distinct ranking options. Users can add other users as friends or subscribe to their videos, add a video as favorite, leave comments and also vote up to five stars according to their level of satisfaction.

YouTube, as well as any online social service that allows users interactions, can be represented by multiple graphs. Using its features one can generate various networks of users based on their social interactions. For instance, considering the users as network nodes, directed edges could link two users if one has added the other as friend, one has left the other user a comment on any of her videos or if both users have added the same video as favorite. We have used in this work a special network of users that, to the best of our knowledge, has never been studied: the *subscription network*.

YouTube subscription network is built up when authenticated users subscribe to videos from other users. Considering the users as network *nodes*, a *directed edge* from user A to user B means that A has subscribed to videos from B. The subscription feature is closely related to the idea of RSS feeds in which users can keep up with their favorite web sites in an automated manner. When the users log in to YouTube they have access to their subscriptions right on the first page. We have opted for the subscription network because analyzing the relations emerged from these explicit social interactions is more suitable to capture the essence of YouTube which is to enable the production, promotion and visualization of *user generated videos*. Even though there may be some professionally generated content, the distribution process is the same. Moreover, it captures a more extensive variety of actions than other networks of users, like the one built up from friendship, for instance.

We designed a web crawler to collect a sample from the subscription network based on a breadth-first search. This technique, also known as *snowball sampling* [14], starts with a set of users and adds all her subscribers and subscriptions to the end of the list of available users. The whole procedure is repeated until the crawler is manually stopped or the entire network component is exhausted. In this work, the top 100 most subscribed users constituted the initial set of users. For each user found in the search, we saved information such as the number of uploads, videos watched and the date he joined the system. Table 1 summarizes the sample collected and show the main network metrics.

Table 1: Summary of YouTube subscription network (CV = Coefficient of Variation).

Crawling period	06-17/Sep/2007
Total number of users	1,467,003
Avg. clustering coefficient	0.07
Avg. reciprocity	0.01
Avg. in-degree (CV)	6.20 (42.88)
Avg. out-degree (CV)	6.20 (14.84)
Avg. uploads per user (CV)	3.41 (5.91)
Avg. video watches per user (CV)	604.61 (3.24)
Avg. channel visits per user (CV)	720.56 (34.84)

4. CLUSTERING METHODOLOGY

This section describes the methodology for clustering the users. We used a clustering algorithm that assigns users to groups through a distance measure that is computed based on the values of a normalized vector representation of the users. The normalized vector representation of the users and the clustering procedure is detailed in the following sections.

4.1 Representing Users as Feature Vectors

We define our user feature vector as a unidimensional vector of length nine, where each position contains an information about the referred user. It is defined as follows:

$user_i = [f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9]$, where the first five positions contain attributes related to each individual user[2] and the last four positions holds attributes stemmed from the user social interactions. Each collected user holds an individual feature vector. The nine features are detailed next.

1. **Number of uploads** (f_1): This is the number of different video clips the user has uploaded to YouYube. It could indicate the potential of the user as a producer.

2. **Number of watches** (f_2): This is the number of different videos the user has watched. It could indicate the potential of the user as a content consumer.

3. **Number of channel views** (f_3): YouTube considers each user as a channel. This is the number of different user information pages, or channels, a user has visited possibly when searching for videos.

4. **System join date** (f_4): Considering February 1^{st}, 2005 as the date YouTube was founded, the system join date is the time elapsed from foundation to the day the user created the account at YouTube.

5. **Age** (f_5): Each YouTube user has also a timestamp relative to the time he performed his last login to the system. We consider as user age the time elapsed between the join date and the last login, or in other words, the amount of time the users had to perform all actions on YouTube.

6. **Clustering coefficient** (f_6): This is a measure[3] of the interconnection between the users and their neighbors. If user A subscribed to videos from user B and also to videos from user C, then there is a high probability of a subscription between users B and C.

7. **Reciprocity** (f_7): Indicates the probability of mutual subscriptions. The reciprocity of user i is the ratio $R_i = A/B$, where A is the number of reciprocal subscriptions user i has made and B represents all user i subscribers and subscriptions.

8. **Out-degree** (f_8): Number of subscriptions made by the user. Also indicates the user potential as consumer.

9. **In-degree** (f_9): Number of subscriptions received by the user. Also indicates the user potential as producer.

These nine features are of different units and, mostly, of different magnitudes. Considering that the distance measure used by the clustering algorithm is calculated based on the value of each feature, we must normalize the data to ensure that the distance is computed with features of equal weight, otherwise the one with the largest scale would prevail over the others. We then normalized the data by the maximum value of each feature so that every feature ranges from 0 to 1.

[2]The number of uploads, watches and channel views of a user represent his/her interactions only in periods when he/she was logged in to the system.

[3]We used Watts and Strogatz definition [21].

4.2 Grouping Users with Similar Behavior

In order to group users that share similar behavioral pattern we have used K-means [9, 13, 15] as the clustering algorithm and the Euclidean distance[4] as the distance measure. Briefly, K-means selects K points in space to be the initial guess of the K centroids[5]. Remaining points are then allocated to the nearest centroid. The whole procedure is repeated until no points switches cluster assignment or a number of iterations is performed.

In addition to the feature vectors, K-means also requires the number of clusters to be created (K) as input. A question then arises: *How many clusters should we choose?* In [16] and [17] the authors suggest that this question can be answered by examining the variation of two metrics: the intracluster distance (average distance between each cluster point and its centroid) and the intercluster distance (average distance between centroids), both characterized by their Coefficient of Variation (CV). The goal is to minimize intracluster CV while maximizing the intercluster CV. The ratio between the intracluster CV and the intercluster CV, denoted by β_{CV}, can help us define the value of K. Varying the number of clusters yields different values for β_{CV}. The best indication for K would be when β_{CV} becomes relatively stable [16, 17].

Once defined the number of clusters, we use the centroids to represent the users assigned to each group. We then analyze the average values of each feature to draw our conclusions and associate each centroid with a different YouTube user behavior.

5. IDENTIFYING USER BEHAVIORS

This section describes the cluster analysis performed on the feature vectors. Figure 1 shows the β_{CV} variation for executions of the K-means with different values of K. One can observe that there is no number of cluster in which β_{CV} *becomes* stable. It is relatively stable for all values of K, meaning that an good value for K is not clear and we must employ a different methodology to select K. The following sections describe the iterative strategy used instead.

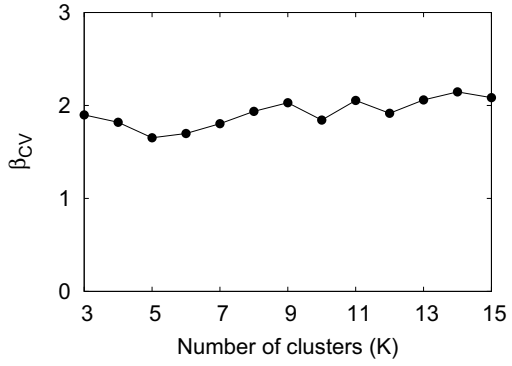

Figure 1: β_{CV} for Varying Values of K

[4]The Euclidean distance is defined as $D = \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2$, where n is the length of user feature vector and x and y are two points in space (two users or one user and the centroid).

[5]The centroid is defined as the point whose coordinates are the average among all points in the cluster.

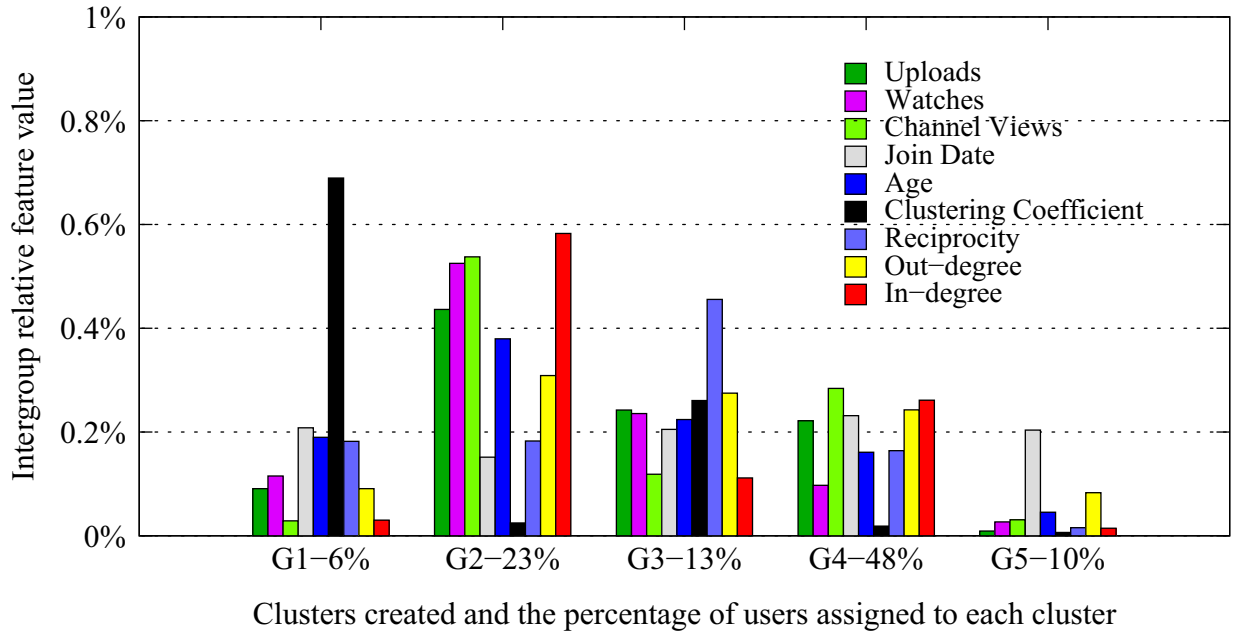

Figure 2: Relative Feature Values for the Five Clusters Created.

5.1 Grouping Users by Iterating over K

We can interpret figure 1 from a different point of view and argue that any value of K is as good as any other, though we must keep in mind that each increment in K yields a new cluster and similarly a new behavior. On the one hand, if we consider a small number of clusters as an aggregation of many possible behaviors, we could expect the most predominant behaviors to be discovered for lower values of K. On the other hand, by selecting a small value for K we may miss some relevant behaviors. Thus, one possible strategy is to analyze the behaviors discovered as K is incremented. However, *up to which value should we increment K?*

Algorithm 1 depicts our methodology. To help us define our stopping criteria, we can take any two centroid vector, for instance C_1 and C_2, and compute the difference between them as $d(C_1, C_2) = abs(\sum_{i=1}^{F}(C_1[i] - C_2[i])/F)$, where $abs(x)$ is the absolute value of x and F is the length of the vector, or the number of features (9, in our analysis).

Algorithm 1 Clustering Identification Algorithm.

```
1: K ⇐ 2;
2: repeat
3:    K ⇐ K + 1;
4:    run K-means algorithm;
5:    for (each cluster k returned by K-means) do
6:       C_k ⇐ centroid of cluster k;
7:       if (∃_{k,x} | d(C_k, C_x) < T) then
8:          merge users from clusters k and x;
9:       end if
10:   end for
11: until (d(C_i, C_n) < T) ∧ (d(C_n, C_j) < T),
          ∃_{i,j,n}, where i ≠ j, j ≠ n, i ≠ n, {i, j, n} ∈ [1, K];
12: manually analyze the features and associate the cluster
    centroids to user behaviors;
```

This difference measure ranges from 0 (equal vectors) to 1 (opposite vectors). If d is below a certain threshold T we can argue that both vectors are very similar and then merge the users of the corresponding clusters into one single larger group. To be conservative, we use a threshold of $T = 10^{-3}$ and define our stopping criteria to be when a newly created cluster merges with an already merged one, for example, $(d(C_1, C_2) < 10^{-3})\ AND\ (d(C_2, C_3) < 10^{-3})$.

Based on the aforementioned methodology, the yielded clusters are displayed in figure 2. On the x axis, each cluster is illustrated by its set of 9 features, along with the percentage of all users assigned to the cluster. On the y axis the intergroup relative feature value, denoted by r_{ij}, measures how a feature i of cluster j is related to the same feature of the other clusters. It is computed as follows $r_{ij} = f_{ij} / \sum_{k=1}^{K} f_{ik}$, where f_{ij} is the value of the feature i of cluster j. This allows us to contrast the same feature i of all clusters. Apart from the absolute values, we can see, for instance, if the values of a feature are evenly distributed between clusters or concentrated in a single one. This can give us valuable hints for the manual behavior inference.

For instance, the join date is roughly evenly distributed among the groups and does not help us to define any behavior. Based on the other features, we describe the five different groups identified:

- **G1** - *Small Community Member*: This group corresponds to users that form small but highly interconnected communities such as family members or colleagues from work or school (small in-/out-degrees and considerably higher clustering coefficient than any other group). These users have low values for their features meaning they are not as active as users with different behaviors. Typically, they create their accounts, subscribe to videos from their communities and then attenuate their craving for interactions.

- **G2** - *Content Producer*: This group corresponds to a typical content producer and 23% of all users. Users are relatively older, meaning they are constantly accessing their accounts. They visit many channels and also watch and upload many videos. As content producers they receive many subscriptions from varying audience (low clustering coefficient). They also subscribe to videos of many users either to be in touch with other producers or as a response for an eventual subscription received (reciprocity). The content produced can be either home made or professional, from users like *BBCWorldwide*, *NationalGeographic*, *universalmusicgroup*, *NBA*, *CBS* or *warnerbrosrecords*.

- **G3** - *Content Consumer*: This group corresponds to a typical content consumer. These users browse through the available videos more than they do with channels. YouTube has a feature that suggests videos related to the one being watched at the moment. They also subscribe to videos more than they receive subscription. Users and their subscription targets seem to share similar interests due to the high clustering coefficient and specially high reciprocity.

- **G4** - *Producer & Consumer*: This group corresponds to users that have both characteristics of producers and consumers. It represents the largest fraction of users (48%). They have moderately large number of subscribers and uploads, like the producers, and also browse many channels to watch and subscribe to videos, like the consumers. Clustering coefficient and reciprocity are low due to varying subscription targets and subscribers audience.

- **G5** - *Other*: A clear identification of this group is not possible because of the low values of every feature. However, possible interpretations, based on the short user age, are that it could correspond either to recent users or users that have abandoned their accounts. The low values of the features corroborate the recency hypothesis. However, when we analyze the absolute values we see that on average these users have subscribed to more videos than they have watched or the number of channels they have visited. One could consider this behavior as suspect in the sense that YouTube keeps a ranking of the most subscribed users. Accounts could have been created just to make a few isolated subscriptions and increase the rank of a few opportunist users that want to raise the popularity of their channels and videos.

In summary, using the clustering methodology, we have identified up to five distinct behaviors in which YouTube users can be classified based on their individual and social attributes. So far, we have managed a total of nine different features, however *do we need all these features to reach the same results?* or, if we consider a smaller set of the most relevant features, *what are the predominant behaviors?* The next section answers these questions.

5.2 Dominant User Behaviors

From the original set of nine features, we have first discarded the system join date due to its roughly equal distribution among the clusters. Considering an arbitrary threshold of 6 months on the user age, we now split the users into two groups in an attempt to distinguish the users that have been on the system for a longer period and then potentially exhibit a more clearly defined behavior. If the clustering procedure executed on both groups do not yield considerably different behaviors, then we can also discard the user age. Users with age above the threshold represent 63.53% of the total and users below the threshold, 36.47%.

The distinction using the age of the users did not bring significantly new insight. Visually inspecting, we note that equivalent behaviors of both groups have similar overall features distribution. A short interval of time seems to be enough for the users to establish their pattern of social interactions and older users typically have larger number of uploads, watches and channel views only because they have been using the system for a longer time than recent users. We have, therefore, also discarded the user age.

The seven remaining features were divided into two different vectors per user. We repeated the same iterative clustering procedure separately, first considering only the three left individual features (f_1, f_2, f_3) and then only the four social features (f_6, f_7, f_8, f_9). In each case, we contrast the profiles found with those described in the last section to verify whether the descriptions would still hold.

Considering only the user individual features is not enough to produce significantly different behaviors. The iterative procedure results in three different groups, however, almost the totality of users (99%) are assigned to one single large cluster. We run the clustering procedure up to K=15 individually for every feature to try to isolate the one, if any, that causes all users to be assigned to one single cluster. Results are similar to when the three features are used jointly. Most of the users have moderate values for all features, meaning they moderately upload, watch videos and visit channels, whereas only a few enthusiasts stand out.

Our next step is to consider only the four social network features. Figure 3 presents the groups obtained. Comparing the clusters described in Figures 2 and 3, we note similarities between clusters S1-G1, S2-(G2,G4,G5) and S3-G3, for the distribution of their feature values are very similar. From the results, we can draw the conclusion: considering only the user social network features we are able to find the most dominant behaviors as we do when we consider all available features. In contrast to the individual features, the social features carry information not only about the individual users, but also implicitly about their neighbors. This aggregated information is stronger and outweighs individual features, thus defining better the user behavioral pattern.

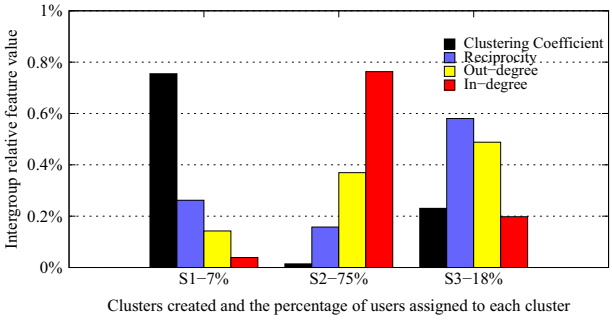

Figure 3: User profiles using social features

Identifying user behavior within an environment in which we know nothing *a priori* is a challenging task and also an empirical process. We use K-means, an unsupervised clustering algorithm, to find the clusters. The algorithm, by definition, runs without any optimization criterion or feedback [9, 13, 15]. Thus, there is no right or wrong number of clusters to find. Defining a *good*, not optimal, number of clusters, and ultimately user behaviors as in our analysis, is part of the problem and closely depends on the nature of the network of users and the data available.

Finally, based on four simple and local networking measures, one can use the methodology we present to assign users to groups with similar behavior and then incorporate this knowledge to improve the systems. The methodology used in this work is not restricted to YouTube. Indeed, it is applicable to online social networking systems in general.

6. CONCLUSIONS AND FUTURE WORK

In current Web 2.0 services, users with different aspirations and goals interact with the site and with other users through a series of multiple interfaces. Different interaction patterns can be observed for different groups of users, and yet specialists must plan accurately their system resource capacities in order to efficiently serve these users. This work presented a methodology for characterizing and identifying user profiles in online social networks.

We crawled data from YouTube and used a clustering algorithm to group users that share similar behavioral pattern. We have identified and discussed up to five different groups in which YouTube users can be classified. In an attempt to verify which attributes are relevant to the clustering procedure we run experiments based on attributes related to each individual user and also based on those stemmed from social interactions. We have shown that attributes stemming from the user social interactions, in contrast to attributes relative to each individual user, are good discriminators and allow the identification of the dominant user behaviors.

Identifying different user behaviors has the potential to improve business and resource management in online social networks. For future directions we could investigate, for instance, recommendation systems that exploit the user behavior to display more apropriate advertisements. We could also exploit the user behaviors to define different classes and develop more accurate performance models for the services.

7. REFERENCES

[1] E. Adar and B. Huberman. Free Riding on Gnutella. *First Monday*, 5(10), 2000.

[2] E. Agichtein, E. Brill, and S. Dumais. Improving Web Search Ranking by Incorporating User Behavior Information. In *Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2006.

[3] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding High Quality Content in Social Media, with an Application to Community-Based Question Answering. In *Proc. ACM Web Search and Data Mining (WSDM)*, Stanford, CA, USA, Feb 2008.

[4] Y.-Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong. Analysis of Topological Characteristics of Huge Online Social Networking Services. In *Proc. Intl. World Wide Web Conference (WWW)*, Banff, Alberta, Canada, May 2007.

[5] Alexa Web Search. `http://www.alexa.com`, 2008.

[6] L. Backstrom, R. Kumar, C. Marlow, J. Novak, and A. Tomkins. Preferential Behavior in Online Groups. In *Proc. ACM Web Search and Data Mining (WSDM)*, Stanford, CA, USA, Feb 2008.

[7] Business Intelligence Lowdown: Top 10 Largest Databases in the World, Feb 2007.

[8] 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. ACM Internet Measurement Conference (IMC)*, San Diego, CA, USA, Oct 2007.

[9] B. Everitt. *Cluster Analysis*. Halsted Press, NY, 1980.

[10] M. Feldman, C. Papadimitriou, J. Chuang, and I. Stoica. Free-riding and Whitewashing in Peer-to-Peer Systems. *IEEE Journal on Selected Areas in Communications*, 24(5):1010–1019, 2006.

[11] D. Fisher, M. Smith, and H. Welser. You Are Who You Talk To: Detecting Roles in Usenet Newsgroups. In *Proc. Hawaii International Conference on System Sciences (HICSS)*, Jan 2006.

[12] P. Gill, M. Arlitt, Z. Li, and A. Mahanti. Youtube Traffic Characterization: A View From the Edge. In *Proc. ACM Internet Measurement Conference (IMC)*, San Diego, CA, USA, Oct 2007.

[13] A. Jain, M. Murty, and P. Flynn. Data Clustering: A Review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[14] S. H. Lee, P.-J. Kim, and H. Jeong. Statistical Properties of Sampled Networks. *Physical Review E*, 73:016102, 2006.

[15] J. Macqueen. Some Methods of Classification and Analysis of Multivariate Observations. In *Proc. Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[16] D. Menascé and V. Almeida. *Scaling for E-Business: Technologies, Models, Performance, and Capacity Planning*. Upper Saddle River, Prentice Hall, NJ, 2000.

[17] D. Menascé, V. Almeida, R. Fonseca, and M. Mendes. A Methodology for Workload Characterization of E-Commerce Sites. In *Proc. ACM Conference on Electronic Commerce (EC)*, pages 119–128, 1999.

[18] D. Menascé, V. Almeida, R. Fonseca, and M. Mendes. Business-Oriented Resource Management Policies for E-Commerce Servers. *Performance Evaluation*, 42(2-3):223–239, 2000.

[19] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and Analysis of Online Social Networks. In *Proc. ACM Internet Measurement Conference (IMC)*, San Diego, CA, USA, Oct 2007.

[20] H. Neto, J. Almeida, L. Rocha, W. Meira, P. Guerra, and V. Almeida. A Characterization of Broadband User Behavior and Their E-Business Activities. *ACM SIGMETRICS Performance Evaluation Review*, 32(3):3–13, 2004.

[21] D. Watts and S. Strogatz. Collective Dynamics of 'Small-World' Networks. *Nature*, 393:440–442, 1998.