Exploring the User-Generated Content (UGC) Uploading Behavior on YouTube

Jaimie Y. Park
Division of Web Science and Technology, KAIST
Daejeon, Korea
jaimie@islab.kaist.ac.kr

Jiyeon Jang
Department of Computer Science, KAIST
Daejeon, Korea
jijjang@kaist.ac.kr

Alejandro Jaimes*
Yahoo Labs
New York City, USA
ajaimes@yahoo-inc.com

Chin-Wan Chung
Dept of Computer Science
Div of Web Science and Tech, KAIST
Daejeon, Korea
chungcw@kaist.edu

Sung-Hyon Myaeng
Dept of Computer Science
Div of Web Science and Tech, KAIST
Daejeon, Korea
myaeng@kaist.ac.kr

ABSTRACT
YouTube is the world’s largest video sharing platform where both professional and non-professional users participate in creating, uploading, and viewing content. In this work, we analyze content in the music category created by the non-professionals, which we refer to as user-generated content (UGC). Non-professional users frequently upload content (UGC) that are parodies, remakes, or covers of the music videos uploaded by professionals, namely the official record labels. Along with the success of official music videos on YouTube, we find the increased participation of users in creating the UGC related to the music videos. In this study, we characterize the UGC uploading behavior in terms of what, where, and when. Furthermore, we measure the relationship between the popularity of the original content and creation of the related UGC. We find that the UGC uploading behavior is different depending on the types of the UGC and across different genres of music videos. We also find that UGC sharing is a highly global activity; popular UGC is created from all over the world despite the fact that the popular music videos originate from a very limited number of locations. Our findings imply that utilizing the information on re-created UGC is important in order to understand and to predict the popularity of the original content.

Categories and Subject Descriptors
H.3.5 [Online Information Services]: Web-based services

*This work was performed while the author was a visiting professor at KAIST, Division of Web Science and Technology, under the WCU (World Class University) program.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author’s site if the Material is used in electronic media.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2576945.

General Terms
Human Factors, Measurement

Keywords
User Generated Content, YouTube, Popularity Analysis, Digital Media, Social Media, Video Streaming Service

1. INTRODUCTION
YouTube is the largest video sharing site in the world. It has been reported that more than 1 billion unique users visit the site, watching over 6 billion hours of videos on a daily basis. YouTube is unique in that it is a truly participatory platform, where many users are self-publishing consumers. YouTube statistics. retrieved on Feb 2nd, 2014. http://www.youtube.com/yt/press/statistics.html.

makes, and covers flourished on YouTube. Inspired by this phenomenon, we study whether the popularity of a music video on YouTube is related to the volume of related UGC. Furthermore, we explore what types of UGC are being uploaded on YouTube, where, and when.

To the best of our knowledge, this is the first work that examines the phenomenon of creating music-related UGC with respect to the success of its original video. The questions addressed in this study include, but are not limited to, the following:

- What types of music video-related UGC do users create?
- Do people create different types of UGC depending on the genre of the music?
- Which countries are involved in uploading UGC? How geographically diverse is the UGC uploading behavior?
- When are UGC uploaded? How long does it take for UGC to be uploaded after the release of original music videos?
- Is there a difference in the UGC upload time depending on the type (parodies, remakes, covers, etc.) of the UGC?
- Is UGC uploading related to the popularity of music videos on YouTube?

The results of our analysis show that the music video popularity and UGC upload frequency are strongly correlated, highlighting the importance of analyzing the UGC uploading behavior in understanding the popularity of the original content. We further find that UGC uploading behavior is different depending on the types of the UGC and across different genres of music videos. We also find that UGC sharing is a highly global activity; popular UGC is created all over the world despite the fact that the popular music videos originate from a very limited number of locations.

2. RELATED WORK

Studies on user-generated content online mainly fall into one of two bodies of research: characterizing the popularity of videos and predicting the popularity of videos. Cha et al.'s work [5] was one of the first works to collect and analyze large-scale YouTube data in order to understand the popularity lifecycle of videos. Following their work, there have been ongoing interests in investigating the temporal dynamics and evolution of video popularity among the research community as well as industry. Related studies mainly focus on modeling, classifying, and predicting the popularity of videos on YouTube [3, 6, 8, 9, 11]. Brodersen et al. [4] address the problem of analyzing the spatial properties in addition to the temporal properties of the videos. Baek and Shin [2] examine the relationship between cultural distance and intercultural consumption of content on YouTube. Susarla et al. [10] study the impact of social influence in popularity of content on YouTube. Our investigation is different from the existing body of work since we study the particular phenomenon where non-professional users “re”create the works of professionals. Although extensive efforts have been put into analyzing the popularity of videos on YouTube, there has not yet been an attempt to understand the behavior of recreating the content in forms of parodies, remakes, or covers. Since these recreated content can potentially influence the popularity of the original content, it is worthwhile to study this new facet of YouTube video popularity and sharing phenomenon.

3. THE YOUTUBE DATA

In this section, we describe the data used for analysis, and the method to collect the data.

3.1 Top Music Videos (MV)

We first collected the 1,000 most-viewed YouTube videos using the YouTube API. The YouTube API returns up to 1,000 most viewed videos across all categories available on YouTube. After the collection we selected the videos that belong to the music category. In total 461 out of the 1,000 most-viewed videos belong to the music category. The majority of them are official videos uploaded by the entertainment agencies or record labels, for many agencies manage accounts on YouTube in order to share content, collect opinions from the audience, and connect with fans. Official videos are easily distinguishable from the non-official videos, namely UGC, because the former are tagged as premium videos by the YouTube service. The official videos were mainly newly released music videos, while some of them were clips of live performances or concert scenes. We chose the top 100 official “music videos” uploaded by the official agencies.

3.2 Top MV-Related UGC

We further collected the user-generated content that is related to the top music videos. Before collecting the UGC, we manually parsed out the title of the song and the name of the artist from the top 100 music video titles. We then used the YouTube API to search for and retrieve the UGC whose titles contain the song title, artist name, and any of the following designated terms: acoustic, cover, dance, flash mob, parody, reaction, remake, and remix. These terms, which represent the popular types of UGC being frequently uploaded by YouTube users, were selected through the following procedure: we collected the top 1,000 most-viewed music-related UGC uploaded by non-professionals, computed the term frequency for every word that appears in at least one of the 1,000 UGC titles, then picked the terms that refer to the type of UGC. Figure 1 is an illustration of the title term frequency for every word that appears in at least one of the 1,000 most-viewed videos. The most-viewed UGC contains the song title, artist name, and any of the following designated terms: acoustic, cover, dance, flash mob, parody, reaction, remake, and remix.

Figure 1: Title Term Frequency Word Cloud for Most-Viewed UGC
frequency in the form of tag cloud. Dance, video, official, cover, music, and live appear to be some of the most frequently used terms in the title to describe the most popular UGC. We also find the terms parody, cover, live, acoustic, flash mob, and concert often being part of the video titles. Using the YouTube API, we retrieved the UGC whose titles are composed of a <song title>, <artist name>, and <designated term>. For example, the top most viewed music video on YouTube is Psy’s “Gangnam Style”. We run YouTube searches for 8 different keywords: <Psy Gangnam Style Acoustic>, <Psy Gangnam Style Cover>, <Psy Gangnam Style Dance>, …, and <Psy Gangnam Style Remix>. Each search returns up to 1,000 results, which leaves us with a maximum of 8,000 UGC for each music video. In the end, we collect 65,628 related UGC for the top 100 music videos.

3.3 Video Attributes

Each video is associated with the following attributes: title, description, uploaded time, username of the uploader, location of the uploader, and the total view count. These attributes can be automatically extracted using the YouTube API. We also tagged the genre of the music for the top 100 most viewed original videos. Music genre was automatically retrieved from the Wikipedia page for each song, where one song may belong to multiple genres.

4. CHARACTERIZING UGC UPLOADING BEHAVIOR ON YOUTUBE

We study three aspects of UGC uploading behavior on YouTube: what kind of UGC related to popular music videos are uploaded, where, and when.

4.1 What

First, we examine the types of UGC and their distribution depending on the genre of the original music videos. Note that these UGC that we analyzed are related to the topmost viewed music videos. Figure 3 shows the frequency distribution of UGC types, for top 1, top 10, and top 100 music videos. Overall, cover, remix, acoustic, and dances are prevalent, while reaction and flash mob types occupy only a small portion of the sample. Figure 2 shows the distribution of UGC types depending on the genre of the music it is dealing with. We find that irrespective of the genre, cover is the most popular type of UGC. Remix UGC is relatively prevalent for dance and electro-pop music, but not as popular for other music genre. As expected, R&B music videos spawned the least amount of the dance UGC and the rap music videos are associated with only a few acoustic UGC. The analysis results show that the UGC creation and uploading behavior varies across different genres, which indicates that UGC type analysis can be considered as a possible feature in identifying the genre of the original music video. In the following subsections, we analyze the temporal and spatial properties associated with UGC across different types and genres.

4.2 Where

We now study the geographical diversity of UGC creating activity on YouTube. We start by describing the locations of top 100 most viewed original videos. We regard the location of a video uploader as the location of the music video, which is given in the standard two-letter ISO 3166 country code format. The top 100 music videos are generated from 6 different countries, United States (87), Great Britain (8),
average numbers are shown in Table 2 in the second column. We now examine the music videos on the top-100 list individually, and compute the number of distinct countries that participated in creating the UGC based on each video. For example Psy’s “Gangnam Style”, the top most popular music video on YouTube, caused participation from 111 different countries. The second most popular music video, Justin Bieber’s “Baby”, led to creation of UGC from 92 different countries. We group the videos based on their global fame, measured by its view count, and the video's “global” fame, measured by the number of UGCs. We find that the relationship between the two is statistically significant (p<.01) but its effect is weak (Spearman’s ρ=.27).

4.3 When

In this section, we look into when the UGC are uploaded after their original videos were uploaded. We first investigate how long it takes for UGC to be uploaded. We computed the time difference between the upload time of the original videos and that of the UGC. For all top 100 original videos, it takes more than a year on average (491 days). For a more detailed analysis, we investigate whether there is a difference in the UGC upload time depending on the types of UGC (Figure 5(a)) and on the genres of the original music videos (Figure 5(b)). Figure 5(a) shows that the reaction UGC takes the least amount of time. Reaction videos are the recent trend on YouTube, where the users film themselves watching a music video and express their feelings and comments while watching it. Since most reaction videos are intended to capture the first impressions, they can be created immediately after the original video is uploaded and do

Table 1: Top 5 Frequent (LOC_original, LOC_ugc) Pairs

<table>
<thead>
<tr>
<th>Country (MV Count)</th>
<th>LOC_original</th>
<th>LOC_ugc</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States (87)</td>
<td>United States</td>
<td>51</td>
<td>194</td>
</tr>
<tr>
<td>United Kingdom (8)</td>
<td>United Kingdom</td>
<td>37</td>
<td>123</td>
</tr>
<tr>
<td>South Korea (1)</td>
<td>United States</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>Australia (1)</td>
<td>United States</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>Spain (1)</td>
<td>United States</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>Italy (1)</td>
<td>United States</td>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td><strong>Total (99)</strong></td>
<td><strong>54</strong></td>
<td><strong>208</strong></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Number of LOC_original

<table>
<thead>
<tr>
<th>LOC_original (MV Count)</th>
<th>Average No. of LOC_ugc</th>
<th>Total No. of LOC_ugc</th>
</tr>
</thead>
<tbody>
<tr>
<td>United States (87)</td>
<td>51</td>
<td>194</td>
</tr>
<tr>
<td>United Kingdom (8)</td>
<td>37</td>
<td>123</td>
</tr>
<tr>
<td>South Korea (1)</td>
<td>110</td>
<td>110</td>
</tr>
<tr>
<td>Australia (1)</td>
<td>98</td>
<td>98</td>
</tr>
<tr>
<td>Spain (1)</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>Italy (1)</td>
<td>73</td>
<td>73</td>
</tr>
<tr>
<td><strong>Total (99)</strong></td>
<td><strong>54</strong></td>
<td><strong>208</strong></td>
</tr>
</tbody>
</table>

Figure 5: Elapsed Time between Original MV Upload and Related UGC Upload (in Days)
not require any preparation in advance. However, the other types of UGC require some time for the users to familiarize themselves with the original music videos such as learning the dance moves or rearranging the songs and instrumentals. Figure 5(b) shows the reggae music UGC takes the least amount of time but have high variability. Also, the UGC of dance music videos is uploaded relatively quickly.

We further study when the UGC is spread out to other countries to understand when and where the UGC is uploaded. Table 3 shows the top 5 quickest uploaded non-local UGC. The fastest upload of UGC occurred within a week from the day the original was uploaded. The UGC was uploaded from Isle of Man after the original video from the United Kingdom was uploaded. We posit that this is due to the cultural and spatial proximity between the two countries, although further study is necessary to validate the hypothesis. Table 4 shows the locations of the top 100 original music videos and the average time it took for the UGC to be uploaded. We find that it takes the longest on average for the US music videos. It may be due to the uneven data distribution since the US has the most original videos and UGC. A notable point is the relatively small time difference for the South Korean music video, “Gangnam Style”. It shows that the UGC on “Gangnam Style” was created in more than 100 countries within four months after the official “Gangnam Style” music video was uploaded.

Table 3: Top 5 Quickest \( (LOC_{\text{original}}, LOC_{\text{UGC}}) \) Pairs

<table>
<thead>
<tr>
<th>( LOC_{\text{original}} )</th>
<th>( LOC_{\text{UGC}} )</th>
<th>Elapsed Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>United Kingdom</td>
<td>Isle of Man</td>
<td>7</td>
</tr>
<tr>
<td>United States</td>
<td>Zambia</td>
<td>9</td>
</tr>
<tr>
<td>United States</td>
<td>Chad</td>
<td>17</td>
</tr>
<tr>
<td>South Korea</td>
<td>Cayman Islands</td>
<td>20</td>
</tr>
<tr>
<td>South Korea</td>
<td>Liechtenstein</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 4: Average Time between Original MV Upload and Related UGC Upload (in Days)

<table>
<thead>
<tr>
<th>( LOC_{\text{original}} )</th>
<th>Avg. Elapsed Days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spain</td>
<td>741</td>
</tr>
<tr>
<td>United States</td>
<td>524</td>
</tr>
<tr>
<td>Australia</td>
<td>485</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>451</td>
</tr>
<tr>
<td>Italy</td>
<td>114</td>
</tr>
<tr>
<td>South Korea</td>
<td>111</td>
</tr>
</tbody>
</table>

5. YOUTUBE VIDEO POPULARITY AND UGC UPLOADING BEHAVIOR

In this section, we explore the relationship between the popularity of an official music video with respect to the volume of related UGC being uploaded on YouTube. As explained earlier, “related” UGC refers to the recreated content of the original music video including acoustics, covers, dances, flash mobs, parodies, reactions, remakes, and remixes.

We examine the top most viewed music video on YouTube, “Gangnam Style” by a Korean pop-star Psy. In particular, we question whether there is a correlation between the view count of Gangnam Style music video and the number of UGC related to Gangnam Style. Figure 7 shows the change in the number of view counts of the official “Gangnam Style” music video (a), and the change in the number of “Gangnam Style” UGC being created (b) over time. The original “Gangnam Style” music video has reached its maximum number of views in October 2012 with a sudden jump in the preceding month and a gradual decrease afterwards. We can immediately notice from the graphs that the changes in the number of related UGC uploaded closely resemble the growth pattern of the “Gangnam Style” music video view count; a sudden leap in the number of uploads lead to a peak in September 2012, followed by a gradual decrease afterwards.

Moreover, we observe that the change in the number of UGC uploads precedes the change in the music video view counts, which suggests the potential of utilizing the information from re-created UGC for predicting the popularity of the original content. We apply the econometric technique of bivariate Granger causality analysis [7] to the monthly view counts for “Gangnam Style” MV over Time

Figure 6: Original Music Video View Counts vs. UGC Upload Frequency

Figure 7: Number of Related UGC Uploads over Time

Figure 6: Original Music Video View Counts vs. UGC Upload Frequency
time series of UGC upload volume and original MV view count. Through this analysis, we can test whether the values of time series $X$ will exhibit a statistically significant correlation with the future values of time series $Y$. In other words, $X$ is said to Granger-cause $Y$ if $Y$ can be predicted better by using the histories of both $X$ and $Y$ than using the history of $Y$ alone. The analysis results in a p-value below 0.001, with a lag value of 1 month. It allows us to reject the null hypothesis that UGC upload time series does not predict the popularity (view counts) of the original music video, indicating that UGC upload volume Granger-causes the original video view count.

6. CONCLUSIONS AND FUTURE WORK

In this study we explored the phenomenon in which YouTube users create and upload videos that are the “re-creations” of the popular music videos uploaded by the official record labels or agencies. We identified different types of user-generated content (UGC) including covers, dances, parodies, reactions, flash mobs, etc., and found that the UGC uploading behavior is different across different types of UGC as well as the genres of the original music videos. In addition, we found that sharing of UGC is a global activity, drawing participation from a world-wide audience, and that it is a long-lasting activity, where it takes more than a year on average for UGC to be uploaded.

The second phase of our analysis involved correlation analysis between the volume of related UGC and popularity of the original content. We found that the extent to which UGC are uploaded on YouTube is positively correlated with the popularity of the original video, where popularity is measured by the number of view counts. It implies that the analysis of UGC uploading behavior can help one in understanding the popularity of the original content. We supplemented this hypothesis through a Granger-causality test, where we provided evidence that the changes in the volume of related UGC uploaded on YouTube is predictive of the changes in the popularity of the original music video. As also discussed in Cha et al’s work [5], predicting the future popularity of content is important for service providers in terms of devising effective proxy caching strategies to accelerate service requests, which could ultimately lead to overall satisfaction of users of the service.

Due to the YouTube policy, it was not possible to crawl all the related UGC of the original music videos. Thereby our analysis is based on the popular music videos and the relevant UGC with the highest view counts. Although this could be a possible source of bias in the analysis results, we believe it is minimal because these top popular videos account for majority of the traffic on YouTube [5]. For future work, we plan to collect user ratings and comments along with the view counts as additional indicators of video popularity, and study the relationship between them and the growth of UGC uploads.

7. ACKNOWLEDGMENTS

This work was supported in part by WCU (World Class University) program under the National Research Foundation of Korea and funded by the Ministry of Education, Science and Technology of Korea (No.R31-30007), and in part by the National Research Foundation of Korea grant funded by the Korean government (MSIP) (No. NRF-2009-0081365).

8. REFERENCES