Analyzing the Video Popularity Characteristics of Large-Scale User Generated Content Systems

- Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon
Why the study of UGC (User generated content)

- “bite-size bits for high-speed munching”
  [Wired magazine]
- Hundreds of millions of Internet users are content consumers AND publishers
- UGC is very different from VoD
UGC vs. VoD

- **Decentralization**
  UGC: Unlimited choice of content and the convenience of the Web
  Early days of TV: Same program at the same time

- **Scale**
  15 days in YouTube to produce 120-yr worth of movies in IMDb

- **Publisher**
  1000 uploads over few years vs. 100 movies over 50 years

- **Length**
  30 sec - 5 min vs. 100 min movies in LoveFilm
Understanding the popularity characteristics of UGC

- Estimate the latent demand
  - may exist due to bottlenecks

- A lack of editorial control
  - problems for content aliasing and copyright infringement
Data
Summary of User-Generated Video and Non-UGC Traces

<table>
<thead>
<tr>
<th>Name</th>
<th>Category</th>
<th>Num. videos</th>
<th>Total views</th>
<th>Total length</th>
<th>Data collection period</th>
</tr>
</thead>
<tbody>
<tr>
<td>YouTube</td>
<td>Ent</td>
<td>1,687,506</td>
<td>3,708,600,000</td>
<td>15.2 years</td>
<td>December 28, 2006 (crawled once)</td>
</tr>
<tr>
<td>YouTube</td>
<td>Sci</td>
<td>252,255</td>
<td>539,868,316</td>
<td>1.8 years</td>
<td>January 14-19, 2007 (daily), February 14, March 15, 2007 (once)</td>
</tr>
<tr>
<td>Daum</td>
<td>All</td>
<td>196,037</td>
<td>207,555,622</td>
<td>1.0 year</td>
<td>March 1, 2007 (once)</td>
</tr>
<tr>
<td>YouTube</td>
<td>Pop*</td>
<td>2,091</td>
<td>avg. 31,689</td>
<td>med. 186 sec</td>
<td>January 13 - February 5, 2007 (daily)</td>
</tr>
</tbody>
</table>

*For globally popular videos in YouTube, we show the average number of views and the median length of videos.

<table>
<thead>
<tr>
<th>Name</th>
<th># Videos</th>
<th>Period</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netflix</td>
<td>17,770</td>
<td>Oct 2006</td>
<td>Customer ratings</td>
</tr>
<tr>
<td>Lovefilm</td>
<td>39,447</td>
<td>Jan 2007</td>
<td>Length and director</td>
</tr>
<tr>
<td>Yahoo! Movies</td>
<td>361</td>
<td>2004–2007</td>
<td>Theater gross income</td>
</tr>
</tbody>
</table>
Goal

- UGC Versus NON-UGC
- Popularity Distribution
- Popularity Evolution over Time
- Aliasing and Illegal Uploads
Part 1: UGC versus NON-UGC

- Content Production Patterns
- Content Consumption Patterns
Content Production Patterns

Over 1,000 videos over a few years

Two orders of magnitude

(a) Number of videos per producer

(b) Distribution of video length
Content Production Patterns

Cultural differences may cause Daum uploaders to be more active on Sundays. An off-peak day for YouTube users.
Content Consumption Patterns
(Scale of Pupularity)

Many unpopular videos in UGC

Median
YouTube (182)
Netflix (561)
Yahoo! (3,843,300)

Zero Viewer (1,782)

Netflix: user customer ratings instead (so lower bound on the graph)
Content Consumption Patterns

• User Participation
  - The video popularity and rating
    a. String positive linear relationship for both UGC and non-UGC
       The correlation coefficient, 0.8 (YouTube), 0.87 (Yahoo! Movies)
    b. The level of active user participation: low
       Only 0.22% account do ratings
       Only 0.16% account do comments

• How Content is Found
  - 47% of all videos have incoming links
  - Nevertheless, the total clicks from these links are only 3%
    (External links are not significant)
Part 2: Popularity Distribution

- Pareto Principle
- Statistical Properties
Why analyze the popularity distribution

• Helps us understand the underlying mechanism

• Helps us answer important design questions
  - The scale-free nature of Web requests: Improve search engines and advertising policies
  - The distribution of book sales: Design better online stores and recommendation engines
Pareto Principle

10% of the top popular videos account for nearly 80% of views

Other online VoD systems show smaller skew!
UGC Video Empirical Plot

Straight Line waists (two orders of magnitude) and truncated both end
Most Popular Videos

Fetch-at-most-once behavior

Two types of user populations
FetchOnce: Behavior of fetching each immutable object only once
Power: Behavior of requesting popular videos multiple times
Most Popular Videos

Fetch-at-most-once
The decay in tail gets amplified for larger $R$ and smaller $V$

Number of Videos ($V$), Users ($U$), Average number of request per user ($R$)
Long Tail opportunities in UCG

Two reasons for a decaying tail below

1. The natural shape of the UGC popularity distribution is CURVED
2. Bottlenecks in the system
   (Information filtering or post-filters)
If it is Naturally Curved?

And decaying tail is due to removable bottlenecks

Potential Benefit from removal of bottlenecks

<table>
<thead>
<tr>
<th>Total potential gain</th>
<th>Ent</th>
<th>Sci</th>
<th>Travel</th>
<th>Food</th>
</tr>
</thead>
<tbody>
<tr>
<td>Num. beneficiary videos**</td>
<td>1.2M</td>
<td>240K</td>
<td>5K</td>
<td>400</td>
</tr>
</tbody>
</table>

* Percent increase in total views obtained from removing bottlenecks.
** The number of videos whose views will increase by removing bottlenecks.
Part 3 : Popularity Evolution over Time

- Popularity distribution Versus Age
- Temporal Focus
- Time Evolution of the Most Popular Videos
Popularity Distribution versus Age

Viewers are mildly more interested in new videos in the average requests
Popularity of individual UGC over Time

If a video did not get multiple requests during its first day, it is unlikely that it will get many requests in the future.

The percentage of videos aged up to \( X \) days that had no more than \( V \) views.
Video rank changes over a range of video ages

Young videos change many rank positions very fast,
Old videos have a much smaller rank fluctuation

Some of the old videos increased ranks dramatically

(a) Popularity distribution based on $\Delta r_{\text{ank}}$
Part 4: Aliasing and Illegal Uploads

- Content Aliasing
- Illegal Uploads
Content Aliasing

- Content Aliasing
  - Exist multiple identical or very similar copies for a single popular event

- Aliasing dilute the popularity of the corresponding event.

- Has a direct impact on the design of recommendation and ranking system
The level of popularity dilution

Undiluted, the original video would be ranked much higher

Recruited 51 volunteers
Identified 1,224 aliases, covering 184 out of the 216 videos

More than two orders of magnitude
Number of aliases versus the age differences

Significant aliases appear within one week

Cross-posted over multiple categories received almost 1,000 times more views.
Contribution

- An extensive trace-driven analysis of UGC video popularity distributions
  - Analysis reveals properties about how users of these systems request UGC videos.
  - Investigate whether video popularity can be modeled as a power-law
  - What characteristics of the system influence the shape of the distribution
  - Examine non-stationary properties of the UGC vide popularity
  - Reveals the level of piracy and content duplication