How To Tell Head From Tail in User-Generated Content Corpora

Nishanth Sastry
King's College London

Abstract

This paper asks whether unpopular tail items in user-generated content corpora are important, and how tail items differ from the popular head items. We develop a user-centric characterisation of the tail which shows that although the head receives a disproportionate share of interest, tail items collectively serve a large number of users. "Tail seekers", with more 'like's in the tail than the head, are shown to constitute more than half the user base. We then examine how interests in head and tail items differ. Temporally, head items are found to enjoy a sustained interest, whereas interest in tail items is short lived. Spatially, interest in tail items is more geographically diverse. Finally, from a social angle, interest in unpopular items appears to be more "viral" than non-viral. We discuss implications of these observations for the handling and distribution of user-generated content.

Introduction

This paper looks at corpora of user-generated content from the content distributor's point of view and asks whether there is value in the items at the tail of the popularity distribution. Should resources be expended in *promoting* or *filtering* tail items? How do the unique characteristics of tail content affect the handling and distribution of content? Answering these questions becomes important with the rise of rich-media user-generated content such as videos, which require extensive infrastructure investment from content providers.

There are two major viewpoints on the importance of tail items (Brynjolfsson, Hu, and Smith 2010): The Long Tail Effect predicts that the ability of the Internet to support tail items can satisfy an unmet demand. The Superstar Effect suggests that the Internet makes it easier to discover the "best" and hence reinforces the popularity of popular items.

Research has thus far focused on categories such as books (Brynjolfsson and Smith 2003), recorded music and feature films (Elberse 2008), where the content corpora, though huge, grow relatively slowly on a daily basis, and are extensively promoted by publishing houses, record labels and films studios. In contrast, large numbers of new user-generated content items are generated every day. These are essentially *uncurated* content items, with no expert reviewers, or vast marketing machinery promoting them.

Our first contribution is an empirical study, using data from five years of activity on the video sharing site vimeo, of whether any of the above theories generalise to the new setting of user-generated content. We find that although there is a Long Tail, it is relatively thin. Items at the head of the popularity distribution receive a disproportionate share of accesses and 'likes', consistent with the Superstar Effect.

However, we develop a user-centric characterisation of the tail and find that a surprising number of users are interested in the tail. Thus, limiting the content catalogue to popular items alone, or selectively promoting them can affect a large fraction of the user base. Indeed, "tail seekers", with more 'likes' in tail than head items, constitute roughly 60% of the user base. We conclude that *while the tail is thin in the collective votes it receives, it is heavy in terms of the number of users it serves*. This is our second contribution.

Our third contribution is a study from temporal, spatial and social perspectives, of differences in how interest in head and tail items arise. Temporally, we find, as expected, that interest in tail items tends to be short lived, whereas head items enjoy a sustained period of interest. Spatially, tail items are geographically more diverse than head items, with a greater fraction of their likes coming from countries with no previous interest in the item. Socially, we find that in contrast to the common notion of videos "going viral", most popular videos obtain likes from socially unrelated users. Tail items have more viral than non-viral likes.

We discuss implications of these findings on handling and distributing tail content. For instance: A short interest window means that tail items need to be promoted soon after they are initially liked, than waiting for them to become popular. Predominance of viral likes suggests that social recommendations are likely to work well in the tail. Geographic diversity makes content delivery difficult because locations of future likes do not correspond to the past.

Preliminaries

We begin by describing our data set and formalising the division of content items in the data into head and tail.

Dataset Our data set is obtained from video sharing site vimeo. Vimeo users can vote for videos by 'like'ing them, link to other users, creating a social network, form groups, or follow video uploaders by subscribing to their channels. Using public APIs made available by the site, all the channels and groups on the site were first crawled, obtaining information about users who are subscribers or members, and videos of the channels and groups. From here, additional video and user objects were obtained by snowball sampling, using links between users to obtain further users and videos.

The vimeo data comprises five years of activity, from Feb 16, 2005 (within 3 months of when vimeo was founded) up to Mar 27, 2010. Some statistics are summarised below.

Video statistics		Social graph statistics	
Videos	443,653	Users	207,468
Likes	2,427,802	Directed links	718,457

Defining the head (popular) & tail (unpopular) Unlike traditional items such as books and movies, user-generated content does not have an alternate physical distribution channel. Thus we cannot easily pick an absolute number (the inventory size of the "brick-and-mortar" stores) to define the head. Picking a fixed percentage (say $x\%$) of the total number of videos also becomes equally arbitrary. Therefore, in this paper, we define the head (popular) items as items whose consumption counts are at least one standard deviation above the mean consumption. Items with a smaller consumption count belong to the (unpopular) tail.

How important is the tail?

This section characterises the importance of items in the tail relative to the popular head items. Following (Elberse 2008), we examine two theories, the Long Tail Effect and Superstar Effect, which predict contradictory outcomes. We examine our data to see which of these applies in the case of user-generated content, and explain discrepancies found.

The Long Tail vs the Superstar Effect At the risk of over simplification, there are two main viewpoints on whether tail items merit attention (Brynjolfsson, Hu, and Smith 2010). The first view, popularised as the "Long Tail effect" by Chris Anderson (Anderson 2006), is based on the observation that unlike physical "brick-and-mortar" retailers which are limited by building sizes and the cost involved in stocking and distributing to a geographically diverse set of stores, online retailers can offer a much larger catalogue, including so-called tail items whose individual sales are low enough to be cost ineffective for brick-and-mortar retailers. One of the earliest works in this area (Brynjolfsson and Smith 2003) examined Amazon book sales and showed that the tail items contribute to a significant fraction of Amazon's total sales.

The second view, dubbed the "Superstar" effect, draws on the economics of Superstars (Rosen 1981), and posits that since the Internet makes it *easier* for consumers to discover the true hits (the superstars), and for these items to be delivered to the consumers wherever they are geographically, popular products can become disproportionately profitable over time (Elberse and Oberholzer-Gee 2006).

The two theories can be tested in terms of the measurable predictions they make: The Long Tail Effect posits that the share of the tail in the online world could be significant, whereas the Superstar Effect places most of the mass of the popularity distribution in the head items.

To test this, we look at the fraction of accesses that would be satisfied if only the top videos (by likes or plays) are available to consumers. The distributions for both are similar (Fig. 1a): we see a familiar 80-20 split, with the least popular 80% of videos only contributing $\approx$20% of plays or likes. Thus, in terms of likes or plays, the tail is long but relatively thin, consistent with the "Superstar" argument.

A user-centric characterisation of the tail Because the Superstar Effect predicts a rich-gets-richer self-reinforcment of item popularity, (Elberse 2008) recommends that firms should not invest in tail items but continue to focus on the few expected "hits". We can apply this to user-generated content by limiting the catalogue to the most popular items, or highlighting them (e.g. by featuring in leader boards such as "Most Popular Videos").

However, Fig. 1a also shows that in terms of users, the tail is extremely important. More than 80% of videos need to be included to satisfy 80% of the user base. In other words, most of the users have *some* interest in the tail. We probe deeper, asking what fraction of user likes are in the tail. This maps the *density* of tail likes for users (Fig. 1b). Nearly half the users have 60% or more of their likes in the tail.Just under 30% of users have *all* their likes in the tail.

Together, these results imply that from a user-centric point of view, the tail is extremely important. Although the popular few items at the head receive a disproportionate number of likes, a large number of *users* have some interest in the unpopular majority of tail items. This is unexpected. Indeed, McPhee's theory of exposure (McPhee 1963), which (Elberse 2008) invoked in explaining the Superstar Effect predicts exactly the opposite based on two effects:

Natural monopoly: The items in the head enjoy a natural monopoly over light users, because they would not have heard of the more obscure tail items.

Double jeopardy Tail items are less well known. Further, they are less likely to be appreciated by those who know.

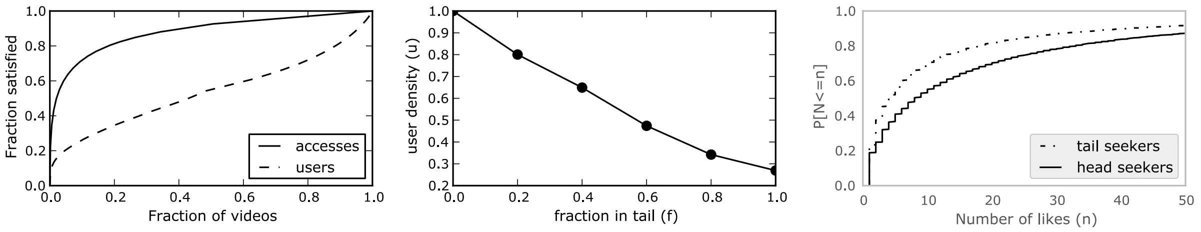
In McPhee's theory, these two effects give rise to a head in which a majority of users, both heavy and light, are interested, and a tail in which not many are interested. This explains the popularity of "hits" and the "Superstar effect".

We find that neither natural monopoly nor double jeopardy hold in the user-generated videos of our dataset. To show this, we first divide users into "tail seekers", users who like more items in the tail than in the head, and "head seekers", who like more head items than tail items.

By definition, head items have a "natural monopoly" over head seekers. If McPhee's theory holds, head seekers should be light users, with fewer likes per capita. Instead, we find that head seekers have a median of 8 likes (mean=25.07) per user, whereas tail seekers are lighter users with a median of 4 likes (mean=22.52). Fig. 1c shows that even in distribution, tail seekers have comparatively fewer likes per capita.

Further, we find that there are slightly *more* ($\approx$60%) tail seekers than head seekers. This negates the double jeopardy prediction, which holds that fewer users would be familiar with or appreciate obscure tail items.

Summary: The characterisation which emerges is of a head and tail enjoyed by nearly equal numbers of users.



(a) (b) (c)

Figure 1: **(a) The tail is heavy in users**. If the tail is defined in terms of number of accesses (plays or likes), the familiar 80-20 rule is obtained with the top 20% of videos satisfying 80% of plays or likes. However, when we look at the fraction of users all of whose likes would be fulfilled by the top x% of videos, more than 80% of videos need to be included to satisfy 80% of the user base. **(b) Likes sets of many users are dense in tail items** $u\%$ of users have a fraction of f or more of their likes in the tail. **(c) Tail seekers are lighter users** The Cumulative Distribution Function (CDF) of the number of likes of tail seekers remains above that of head seekers, indicating that tail seekers are comparatively light users with fewer likes per user.

However, the head as we have defined it constitutes a mere 1.5% of videos. Thus, head items typically obtain more likes than tail items, explaining the thin tail in terms of likes. Figs. 1a–1b show that user interest extends far into the tail. Thus, this picture is robust to other choices for the stopping point of the head and beginning of the tail.

How does the head differ from the tail?

Having established that the tail items are important to the user base, we next identify temporal, spatial and social characteristics in which the tail differs from the head, and study how they can be useful to content providers and distributors.

Temporal perspective: tail items are short lived As corpora of user-generated content continue to grow rapidly, it is natural to ask what is the *effective* catalogue size. Do old items continue to be relevant, or are new ones more important? This can be captured as the interest windows of items (time of last 'like' minus time of first 'like'). Fig. 2a shows, not unexpectedly, that popular stories get accessed over a much longer time window than unpopular stories.Thus, popular stories can see a *sustained* period of interest. In contrast, tail items enjoy only a brief surge of interest.

Based on these results, content providers can decide, for example, to offload only recent items to an expensive content delivery network, or replicate them more widely by other means. Similarly, search results could be made more relevant by attaching greater importance to new tail items, and tail items should be promoted before their interest window expires, rather than waiting for them to become popular.

Spatial perspective: The geographically diverse tail Next, we examine where the likes of each video come from. In Vimeo, each user can enter an arbitrary string description of their location. We use Google and Yahoo geocoding APIs to first map these strings to a latitude-longitude pair, and then obtain the country of the user. We then examine the countries from which 'likes' for videos arise.

Although there are users from 174 countries, one country, United States, accounts for ≈20% of all likes. The top two countries (US and UK) together account for 31%. This skewed usage suggests that overall, a strategy of focusing on the top countries can be useful. For instance, non-specific advertising could focus on ads relevant to the top countries.

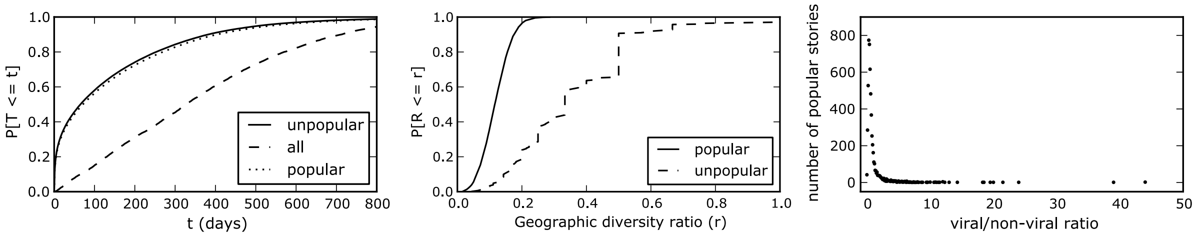
However, on the level of individual videos, tail and head items behave differently. We define geographic diversity as the ratio of the number of distinct countries from which the likes of a video originate, to the total number of its likes. Fig. 2b shows that unpopular videos are more geographically diverse than popular ones. It means that the geographic profile of previous users may not be revealing of future users. This makes it costlier to distribute content such as streaming videos, which need to be replicated close to viewers. It can also cause difficulties for targeted advertising. It can also cause difficulties for targeted advertising.

Social perspective: Impact of viral accesses We next examine how the social network impacts on the popularity. Inspired by theories of viral propagation of awareness about stories, products, etc. that have become prominent recently (e.g. (Gladwell 2002)), we tested the hypothesis that the stories which become popular are the ones that spread successfully within the social network.

Our definitions of viral and non-viral accesses are adopted as follows. If an item is liked by a user after a direct friend on the social network has liked it, we term the access as *viral*. In contrast, if no direct friend has liked the item, then the access is termed as *non-viral*.

Fig. 2c bins videos in the Vimeo trace by the ratio of viral to non-viral likes (rounded to one decimal place). It then measures the number of popular stories. It can be seen that as the ratio of viral to non-viral likes increases, the number of popular stories falls drastically. We conjecture that while there may be individual successes of "viral marketing" strategies, in general, a story which spreads mainly by viral propagation remains localised to one part of the social network. Whereas, an inherently popular story is independently (non-virally) liked in several parts of the social network.

Clearly the ratio of viral to non-viral accesses can be used for multiple purposes, including popularity prediction, as above. Other uses include social recommendations for tail items, which is likely to be more successful when the fraction of viral likes is high.



(a) (b) (c)

Figure 2: **(a) Popular items have sustained period of interest**: CDF of time window of interest shows that unpopular items have briefer windows of interest than popular items. **(b) Interest in unpopular items is geographically diverse**: Geographic diversity ratio measures the fraction of likes of an item that come from a new country. The CDF of this value for popular and unpopular videos is shown. Unpopular videos have a much larger fraction of their likes drawn from a different country. **(c) Non-viral accesses predominate in popular stories**: x-axis shows the ratio of number of viral to non-viral likes. Y axis shows the number of popular stories that have the corresponding ratio. Most popular stories have a low viral to non-viral ratio.

Related Work

A number of studies (Gill and others 2007; Cheng, Dale, and Liu 2008; Cha and others 2009) have looked at various properties of YouTube videos. However, they do not focus on the tail items, which is our goal. Further, although a few studies have looked at the relative popularity of tail items in Video on Demand (Yu and others 2006), and podcasts (Gunawardena and others 2009), very few have developed a user-centric notion of the long tail. (Goel and others. 2010) explores the nature of the tail of user preferences and finds that in non user-generated (curated) content corpora such as movies and music, most users have an interest in and consume items from the tail. Here, we precisely characterise this interest in terms of user density in the tail, and also show that this generalises to new setting of user-generated videos.

Studies of information propagation in Flickr (Cha, Mislove, and Gummadi 2009) and Amazon (Leskovec, Adamic, and Huberman 2007) find evidence that purely viral propagation is largely ineffective. (Lerman and Ghosh 2010) examines the spread of information in digg and twitter and find a predominance of viral accesses similar to us. Our novelty is in identifying differences in viral/non-viral 'like' ratios between head and tail, and suggesting ways to exploit this.

Summary

Examining user-generated videos shared over five years on vimeo.com, this paper first showed that although the head receives a majority of 'like's, user interest extends far into the tail, with many users being "tail seekers" with more 'like's in the tail than the head. Thus, from a user-centric point of view, it is important to effectively distribute and promote tail items. We then discussed factors which help or hinder this: Tail items have a small time window of interest, which helps dramatially reduce numbers. However, interest in the tail is also spatially diverse, which complicates delivery of content such as streaming videos. Finally, likes for tail items are predominantly "viral" which suggests use of techniques such as social recommendations to highlight interesting videos.

We have observed similar results (not shown) with data from social news sharing site, digg. Thus, these conclusions may generalise to other user-generated content.

References

Anderson, C. 2006. *The Long Tail: Why the future of business is selling less of more*. Hyperion Books.

Brynjolfsson, E., and Smith, M. 2003. Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers. *Management Science* 49(11):1580–1596.

Brynjolfsson, E.; Hu, Y.; and Smith, M. 2010. Research commentary—long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns. *Information Systems Research* 21(4):736–747.

Cha, M., et al. 2009. Analyzing the video popularity characteristics of large-scale user generated content systems. *IEEE/ACM Trans. Networking* 17(5):1357–1370.

Cha, M.; Mislove, A.; and Gummadi, K. P. 2009. A measurement-driven analysis of information propagation in the flickr social network. In *Proc. WWW '09*.

Cheng, X.; Dale, C.; and Liu, J. 2008. Statistics and social network of youtube videos. In *IWQoS 2008*, 229–238. IEEE.

Elberse, A., and Oberholzer-Gee, F. 2006. *Superstars and underdogs: An examination of the long tail phenomenon in video sales*. Division of Research, Harvard Business School.

Elberse, A. 2008. Should you invest in the long tail. *Harvard Business Review* 86(7/8):88–96.

Gill, P., et al. 2007. Youtube traffic characterization: a view from the edge. In *Proc. SIGCOMM Internet Measurement Conf. (IMC)*.

Gladwell, M. 2002. *Tipping Point*. Back Bay Books.

Goel, S., and others. 2010. Anatomy of the long tail: ordinary people with extraordinary tastes. In *Proc. WSDM*.

Gunawardena, D., et al. 2009. Characterizing podcast services: publishing, usage, and dissemination. In *Proc. SIGCOMM IMC*.

Lerman, K., and Ghosh, R. 2010. Information contagion: An empirical study of the spread of news on digg and twitter social networks. In *Proc. ICWSM*.

Leskovec, J.; Adamic, L. A.; and Huberman, B. A. 2007. The dynamics of viral marketing. *ACM Trans. Web* 1(1):5.

McPhee, W. 1963. *Formal theories of mass behavior*. Free Press of Glencoe New York.

Rosen, S. 1981. The economics of superstars. *The American Economic Review* 71(5):845–858.

Yu, H., et al. 2006. Understanding user behavior in large-scale video-on-demand systems. *Oper. Syst. Rev.* 40(4):333–344.