The Head and Tail of Wikipedia: the Value of the Long Tail of the Information Warehouse

Bhargav Srinivasa Desikan Institute for Public Policy Research
Akhil Arora EPFL
Martin Gerlach Wikimedia Research
Robert West EPFL

Keywords: Wikipedia, Quasi-experimental, Information and Society, Temporal Data Analysis, Internet Studies

Introduction

Wikipedia, the largest online encyclopedia and platform for open and free knowledge, has long had a debate between the 'deletionists' and 'inclusionists' - those who believe that the millions of articles on Wikipedia that receive less attention should be deleted, versus those who believe that they should be included in the encyclopedia (Lam and Riedl, 2009), (Warncke-Wang et al., 2015). This l̈ong tail" of Wikipedia has also been growing as a consequence of bots, language translation tools, editors and enthusiast adding pages for sports, and culture, as well as efforts by Wikimedia to reduce knowledge gaps. In this work, we use the framing of the head and the long tail, to understand what is the value of the long tail in the context of Wikipedia.

Drawing parallels from the business world, where the 'long tail' concept underpins the success of platforms like Amazon and the appeal of large cities(Anderson and Nissley, 2006), we investigate if a similar benefit exists for Wikipedia. The challenge lies in assessing the value of Wikipedia's long tail, as its costs and benefits are more nuanced than in a commercial context (where it can boil down to finances). Our study aims to quantitatively identifying a consistent 'head' and 'tail' in Wikipedia's various language editions, examining their temporal stability and composition, and attempting to understand what might constitute as value or quality.

Through a quasi-experimental observational study, we compare articles that transition from the tail to the head against those that directly enter the head in different language editions. By using metrics such as network connectivity (in and out degree), edit frequency, and article quality proxies, we evaluate whether time spent in the tail equips an article to better meet information needs when it is required - as it enters the core for the first time. Our rationale behind these metrics is that a well-connected article is better able to meet information needs by providing deeper context, and that an article with more characters, multimedia content, and richer edit history leads to more balanced and higher quality articles.

Having identified a stable head and tail over time, and conducting the quasi-experimental study (described below), we find early, promising results where an article that goes from the tail to the head has higher article quality and integrated network structure than an article that goes straight to the head. These results suggest that a long tail of Wikipedia has an important structural role.

We identify the head and tail of Wikipedia over time by ranking monthly page views of multiple language editions over five years. We then plot ranked page views versus the total number of pages, and identify a hinge or inflection point, where the slope changes such that adding another page does not add proportionally higher aggregate views. We identify the consistency of the head and tail by conducting a sensitivity analysis and find that the range of the cutoff point of the head is between 9 to 12% across all language editions and months that we have chosen. Figure 1 showcases an example of English Wikipedia, where 11.55% of the ranks contribute to 88.14% of pageviews. We also find a consistency in the make-up of the head and tail, with a "canon" of articles that regularly appear in the head, as well as more topical articles related to current trends.

In our quasi-experimental setup, we find pairs of languages with similar size in terms of pages and users. After matching Wikipedia language editions, we then find pairs of pages with the same WikiData ID but as different pages in the language editions. From these pages, we identify pairs of such pages where one article was in the tail for at least one month or longer and moved to the core, versus moving to the core directly after article creation. We showcase early results of this analysis in table 1.

Our findings indicate that articles with a tenure in the tail exhibit higher network connectivity, more edits, and better quality metrics, suggesting that their development in the tail primes them for effectiveness when demand spikes. This, coupled with the argument that a long tail supports diverse interests (among potentially neglected communities) and contributes to an inclusive knowledge base, underscores the structural advantages of maintaining a long tail in Wikipedia.

Future steps would involve exploring other metrics of identifying what makes an article as better able to satisfy

an information need, as well as robustness tests to ensure that the effect we see is indeed from the existence of the article in the tail.

This work makes pivotal contributions by analyzing five years of pageview logs to delineate a stable core and tail in Wikipedia, and by elucidating some of the quantitative and qualitative value of the tail. This not only advances the discussion on deletionists versus inclusionists but also informs strategies for curating and expanding this vital online knowledge repository.

Results

Quality Metric	Avg Metric Val. in Tail to Head	Avg Metric Val. for Direct to Core	T-statistic Value	p-value
Out links	93.5597	77.5147	6.2733	4.4547e-10
In links	67.4000	44.5003	4.2776	1.9919e-05
Image count	36.8311	33.3043	2.9692	0.0030
Total section count	7.5585	7.0519	3.5979	0.0003
Log of character count	8.536	8.4649	3.471	0.00052

Table 1: Quality Metrics Comparison Table

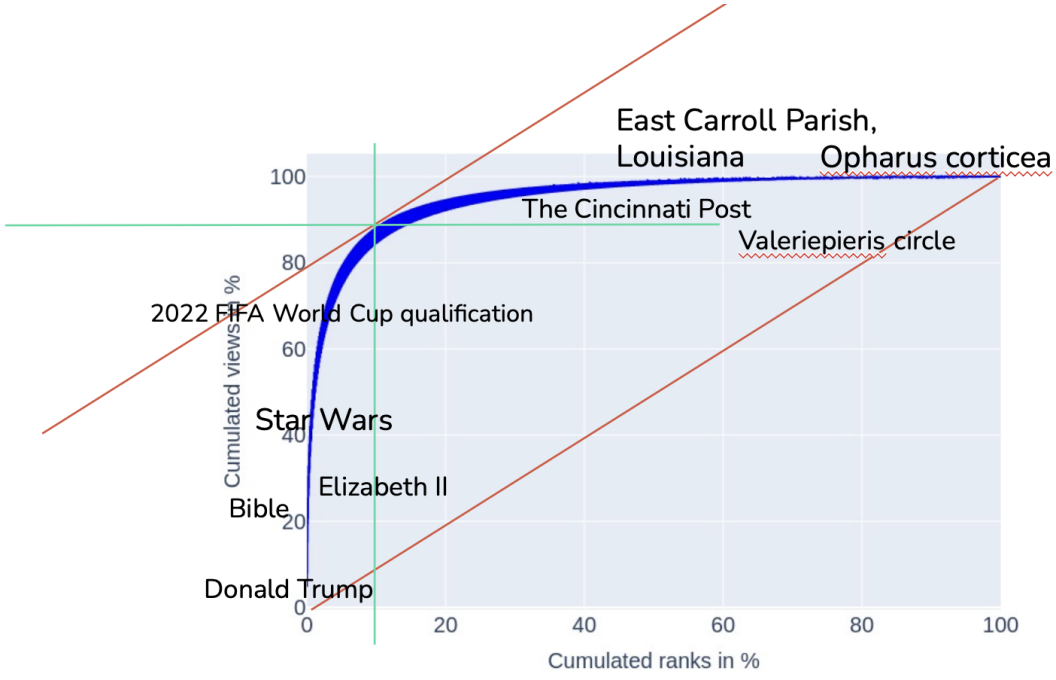

Figure 1: An example of identifying the hinge point for a months aggregate page views for English Wikipedia. Examples of different pages on the curve to illustrate popularity.

References

[Anderson and Nissley2006] Chris Anderson and Christopher Nissley. 2006. The long tail.

[Lam and Riedl2009] Shyong (Tony) K Lam and John Riedl. 2009. Is wikipedia growing a longer tail? In *Proceedings of the 2009 ACM International Conference on Supporting Group Work*, pages 105–114.

[Warncke-Wang et al.2015] Morten Warncke-Wang, Vivek Ranjan, Loren Terveen, and Brent Hecht. 2015. Misalignment between supply and demand of quality content in peer production communities. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 9, pages 493–502.