Orphan Articles: The Dark Matter of Wikipedia

Akhil Arora
EPFL
akhil.arora@epfl.ch

Robert West
EPFL
robert.west@epfl.ch

Martin Gerlach
Wikimedia Foundation
mgerlach@wikimedia.org

Abstract

With 60M articles in more than 300 language versions, Wikipedia is the largest platform for open knowledge. While the available content has been growing continuously at a rate of around 200K new articles each month, very little attention has been paid to the discoverability of the content. One crucial aspect of discoverability is the integration of hyperlinks into the network so the articles are visible to readers navigating Wikipedia. To understand this phenomenon, we conduct the first systematic study of *orphan articles*, which are articles without any incoming links from other Wikipedia articles, across 319 different language versions of Wikipedia. We find that a surprisingly large extent of content, roughly 15% (8.8M) of all articles, is de facto invisible to readers navigating Wikipedia, and thus, rightfully term orphan articles as the *dark matter* of Wikipedia. We also provide causal evidence through a *quasi-experiment* that adding new incoming links to orphans (de-orphanization) leads to a statistically significant increase in their visibility in terms of the number of pageviews. We further highlight the challenges faced by editors for de-orphanizing articles, demonstrate the need to support them, and provide potential solutions for developing automated tools based on cross-lingual approaches. Overall, our work not only unravels a key limitation in the link structure of Wikipedia and quantitatively assesses its impact but also provides a new perspective on the challenges of maintenance associated with content creation at scale in Wikipedia.

Keywords: Wikipedia, knowledge gaps, multilinguality, visibility, quasi-experiment, link recommendation

Introduction

Wikipedia is the largest multi-lingual platform on the Internet for open and freely accessible knowledge. As of November 2022, Wikipedia comprised 60M articles across 319 different language versions, and it has since been growing at a rapid rate of around 200K articles per month. In fact, in order to bridge knowledge gaps (Redi-etal, 2021), there have been a plethora of efforts to systematically add content that is currently absent, e.g., through different initiatives such as organized groups and campaigns, or translating articles across languages (Wulczyn-etal, 2016). As a result, one of the main challenges is how to maintain this ever-increasing volume of content. For example, it is crucial to properly integrate new articles into the existing network structure so that readers can find these articles through hyperlinks, which is one of the main ways to access content on Wikipedia. While the largest share of traffic to Wikipedia comes from search engines, a substantial fraction (38%) of pageviews result from traffic via internal hyperlinks (Piccardi-etal, 2023).

In this work, we explore the question of the lack of visibility of articles in more than 300 language versions of Wikipedia. We specifically focus on so-called *orphan articles*, which are defined as articles that do not have any incoming links from other articles in the main namespace of Wikipedia (`https://w.wiki/6hXb`). These articles are of particular interest since they are de facto invisible for readers navigating hyperlinks in Wikipedia. Specifically, we aim to address the following research questions:

- **RQ1:** What are the key characteristics of orphans?
- **RQ2:** Does adding incoming links (de-orphanization) increase the visibility of orphan articles?
- **RQ3:** What is the current state of de-orphanization and what are the potential ways to improve it?

Data and Methods

We consider 319 different language versions of Wikipedia and collect data spanning 7 monthly snapshots ranging from August 2022 to February 2023. Unless stated otherwise, the results presented in this paper are based on the monthly snapshot of November 2022. For other snapshots, the results portrayed similar trends, and are therefore omitted. We extract the link networks among articles and match links across languages using the articles' Wikidata item ids. In addition, for each article, we extract the following features: topic, quality, time since creation, whether it was created by a bot, whether it is a disambiguation page, the gender (for biography articles). In total, we have 60M articles and 3.5B links.

To answer the aforementioned research questions (RQ1-RQ3), we conduct the first systematic study on orphan articles in Wikipedia and show that orphans make up a surprisingly large fraction of articles. We also establish causal evidence through quasi-experiments that orphan articles are significantly less visible than non-orphan articles. We then describe the challenges faced by editors in addressing this issue and sketch potential solutions to develop models to support their efforts, demonstrating the opportunities for using our insights in future works. Together, these results provide a new perspective on maintenance costs associated with content creation and challenges in making existing knowledge discoverable.

Results

Many orphans. The number of orphans is surprisingly large: 8.8M (14.7%) out of 60M articles do not have any incoming links (Fig. 1). This observation is not limited to only a few or small Wikipedia language versions, rather for more than 100 languages the percentage of orphans is above 30%, including Egyptian Arabic (78%) and Vietnamese (50%), which are among the 20 largest Wikipedia languages. In comparison, the number of dead-end articles, i.e., articles without any outgoing links, is very low across all languages (less than 0.5%). We find that orphan articles are negatively correlated with being: (1) of higher quality and (2) being about the topic of history and society, while possessing a slight positive association with being newer. More importantly, we showed that orphan articles encode structural biases: biography articles about women are substantially more common among orphans than expected from their overall frequency.

Lack of visibility. Orphan articles have, in general, fewer pageviews than non-orphan articles. Additionally, we establish causality via a quasi-experiment (Fig. 2). We consider all orphans that were de-orphanized by editors in a wiki in a given month as our treatment group. In order to rule out potential confounders (such as general increase in interest in the specific topic), we consider a control group comprised of the *same* article in another wiki in which it remained an orphan. Considering the difference-in-difference of the number of pageviews as a proxy for the effective visiblity of each article, we find, on average, a statistically significant overall increase of 6.5% ($p < 10^{-10}$) in the number of pageviews for the treatment group in comparison to the control group (cf. Fig. 3). We also found that this increase is mainly driven by internally-referred pageviews from other Wikipedia articles which contain a link to the de-orphanized article.

Challenges for editors. The rate of organic de-orphanization is alarmingly low (Fig. 4). For the snapshots we considered, editors de-orphanized ~35K orphan articles. While this constitutes an impressive effort by the community, at that rate it would take approximately 20 years to de-orphanize all orphan articles (assuming no newly created orphans). One hypothesis is that existing tools do not support editors in addressing this issue effectively. For example, FindLink (the tool suggested to editors in the orphans maintenance template) generally does not yield many results for orphan articles, especially for smaller languages. However, our results show that an orphan article in one language is not always an orphan in other languages. This suggests that we can develop an approach for identifying articles from which to link to orphans via link translation. Our results shows that this could be effective for 5.5M (62%) orphan articles.

Discussion/Conclusions

Studying orphan articles in more than 300 languages in Wikipedia, we characterized the surprisingly large extent of content that is de-facto invisible for readers navigating hyperlinks in Wikipedia. We proposed a simple and interpretable approach to suggest new incoming links to orphan articles at scale based on link translation, which outperforms existing tools available to editors. This approach can improve the visibility of orphan articles, in line with previous natural experiments demonstrating a spillover effect of attention in Wikipedia (Zhu-etal, 2020). Furthermore, we believe this can help address structural biases such as the gender gap in Wikipedia. For example, visibility of biographies of women is systematically lower than men (Wagner-etal, 2016)). While community-driven campaigns are successful at adding and improving the content about women, they are less successful at addressing structural biases that limit their visibility (Langrock-etal, 2022). Existing link recommendation algorithms are prone to reinforcing those biases and can reduce the visibility of minorities (Ferrara-etal, 2022).

References

[Ferrara-etal2022] Ferrara-etal. 2022. Link recommendations: Their impact on network structure and minorities. In *WebSci*.

[Langrock-etal2022] Langrock-etal. 2022. The Gender Divide in Wikipedia: Quantifying and Assessing the Impact of Two Feminist Interventions. *Journal of Comm.*, 72(3), 02.

[Piccardi-etal2023] Piccardi-etal. 2023. A Large-Scale characterization of how readers browse wikipedia. *ACM TWEB*.

[Redi-etal2021] Redi-etal. 2021. A taxonomy of knowledge gaps for wikimedia projects (second draft).

[Wagner-etal2016] Wagner-etal. 2016. Women through the glass ceiling: gender asymmetries in wikipedia. *EPJ Data Science*, 5(1).

[Wulczyn-etal2016] Wulczyn-etal. 2016. Growing wikipedia across languages via recommendation. In *WWW*.

[Zhu-etal2020] Zhu-etal. 2020. Content growth and attention contagion in information networks: Addressing information poverty on wikipedia. *Inf. Systems Research*, 31(2):491–509.

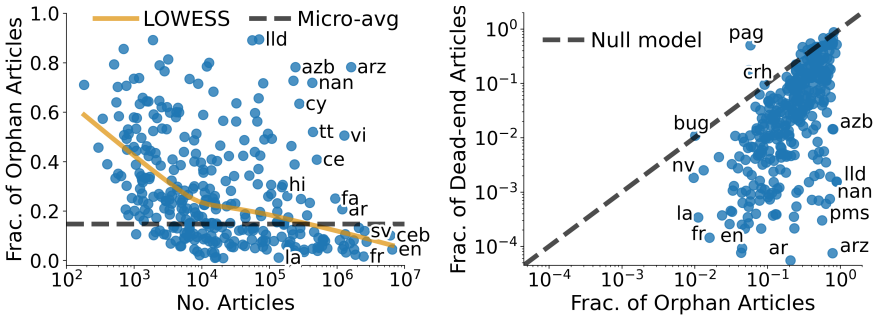


Figure 1: (Left) Analyzing the extent of orphan articles across all Wikipedia language versions. (Right) Comparing the extent of orphans with that of of dead-end articles across all Wikipedia language versions.

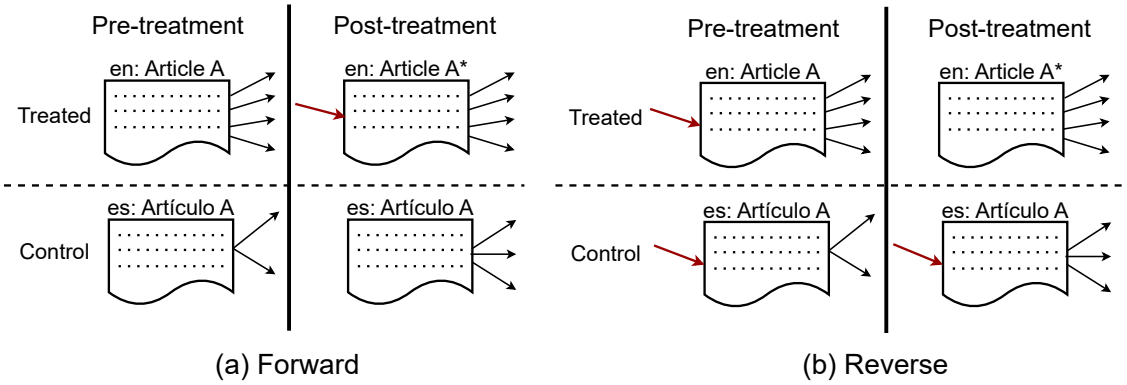


Figure 2: A pictorial representation of the quasi-experiment: (a) Forward: an article that receives a new incoming link (denoted in red font) is considered as treated, whereas the same article in another language that does not receive any new incoming links is considered as control; (b) Reverse: an article that loses an incoming link is considered as treated, whereas the same article in another language that does not lose any incoming links is considered as control.

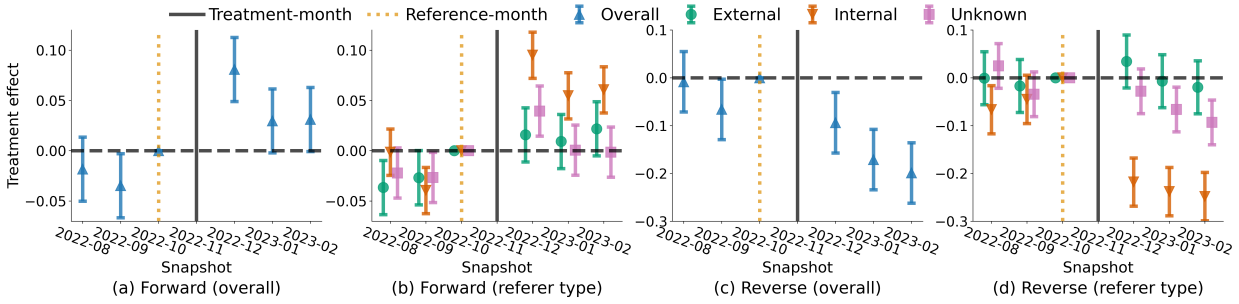
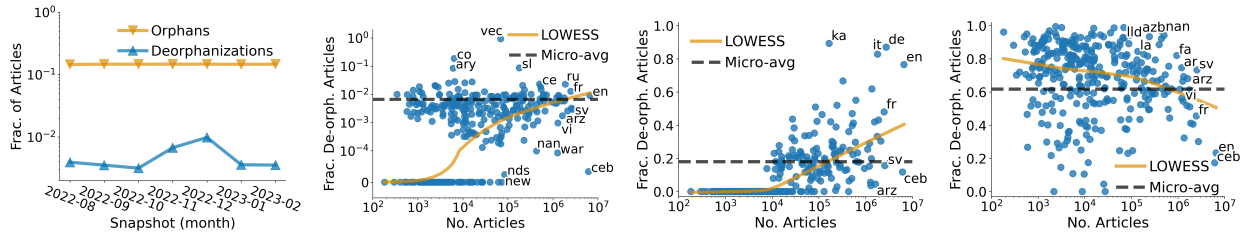


Figure 3: Per-month DiD treatment effect with 95% CIs for the (a)-(b) forward and (c)-(d) reverse setup considering November 2022 as the treatment month.



(a) Orph. vs. De-orphanization (b) De-orphanizations per Wiki (c) Findlink (d) Cross-lingual

Figure 4: Analyzing (a)-(b) the current state of de-orphanization, and the fraction of orphans that can be potentially de-orphanized using (c) Findlink, and (d) Cross-lingual approaches across all Wikipedia language versions.