Orphan articles: The 'dark matter' of Wikipedia

May 17 2024, by Tanya Petersen

Wikipedia logo.

Wikipedia is the largest platform for open and freely accessible knowledge online yet, in a new study, EPFL researchers have found that around 15% of the content is effectively invisible to readers browsing within Wikipedia. They have developed a new tool to help overcome this. The work is published on the *arXiv* preprint server.

With 60 million articles in more than 300 language versions, Wikipedia's available content grows continuously at a rate of around 200 thousand new articles each month. Readers often discover new knowledge and dig deeper into a subject by clicking hyperlinks that connect one article to the next. But what about Wikipedia articles that no other articles link to?

These are commonly referred to as 'orphan' articles and to better understand this phenomenon EPFL researchers from the Data Science Laboratory (DLAB) in the School of Computer and Communication Sciences, in collaboration with the Research Team at the Wikimedia Foundation, conducted the first systematic study of orphan articles across all 319 different language versions of Wikipedia that existed at the time the study was conducted.

"Wikipedia is a network just like roads, the internet, chemical compounds, or genes, and any network has a basic concept of navigability so you can go from one place to another. Information networks are organized in particular hierarchies and we were curious to understand articles that were not reached by anyone. That's how we started to look into orphan articles," explained Akhil Arora, a Ph.D. researcher in DLAB and lead author of the study "Orphan Articles: The Dark Matter of Wikipedia."

The researchers found that almost 9 million articles on Wikipedia across all languages—around 15%—were orphans, effectively invisible to readers browsing within Wikipedia, existing across nearly all topic areas on the platform. In general, pageviews received by non-orphan articles are twice as many as the pageviews of orphan articles. Beyond simple correlations, the researchers also established a cause-and-effect relationship between the addition of in-links to orphan articles and an increase in their pageviews.

The lack of visibility of orphan articles comes down to the way users search and view pages on Wikipedia. The first is via a search engine, where a user is pointed to a particular Wikipedia page; the second is while using Wikipedia as an encyclopedia and clicking through from one article to another and the third is a combination of both.

In all these scenarios, an editor will not only need to add links in the outgoing direction from the article they are editing but will need to know all the relevant Wikipedia articles that could potentially link inwards, and this is a difficult prospect.

"An editor is editing something they know a lot about so they are able to add outward links to other articles," said Arora. "Reversing directionality introduces so many difficulties because they may not be an expert on other topics and articles; sometimes these relationships are not symmetrical and the universe is the entirety of Wikipedia."

The research found that there are large discrepancies across languages. In more than 100 languages, the percentage of orphan articles is more than 30%, with a particularly high figure for Egyptian Arabic (78%) and Vietnamese (50%). Both are among the 20 largest Wikipedia language versions. This points to the challenge of a lack of editor capacity in some languages and demonstrates the need to improve existing tools, such as FindLink, that support editors in this task.

One interesting finding of the study is that an orphan article in one language is not always an orphan in other languages and this led the researchers to develop a new approach for identifying articles from which to link to orphans via link translation.

"If the same article is not an orphan in another language, it means the editors in that community were able to find other articles that could link to this article. So we simply just transferred the link from other languages to the language in which the article was an orphan. We found this approach was able to suggest links for more than 63% of the orphan articles," said Arora.

The EPFL team is continuing to collaborate with researchers at the Wikimedia Foundation on ways this approach could be made available as a tool (see the initial prototype) to improve the experience of readers on Wikipedia. It is also using AI to help this effort on two fronts.

First, the researchers are working on graph neural networks to organize link recommendations that will serve as a basis for the tool. Second, similar to a heat map, they are developing an additional tool that can guide editors as to where in a page text they should consider adding new concepts that will then use generative AI to suggest some starting text.

Importantly, volunteer editors improve, edit, and audit the work done by AI. The approach to AI on Wikipedia has always been through "closed loop" systems, in which humans are in the loop.

"The editor community is doing its service to the world but there are not enough of them, particularly in smaller languages. One of our goals is to better support editors because it can be a daunting task to write and maintain articles. Wikipedia is an incredible open access service and this is why the tools that we're building are so helpful to editors doing this valuable work," concluded Arora.

More information: Akhil Arora et al, Orphan Articles: The Dark Matter of Wikipedia, *arXiv* (2023). DOI: 10.48550/arxiv.2306.03940