

The Webonization of Wikipedia: Characterizing Wikipedia Linking Across the Web

Keywords: Wikipedia, World Wide Web, Wikipedia’s Web presence, Linking patterns, Common Crawl

Extended Abstract

Introduction. Wikipedia’s status as the de-facto encyclopedia on the Internet has rendered it one of the most visited websites in the world. This unique status and prosperity has further inspired rich bodies of academic literature on Wikipedia, providing insights about its structure, readership, knowledge gaps, as well as user behaviours. Initial headways have also been made in assessing Wikipedia’s role beyond Wikipedia through an analysis of its impact on law and science. But no research has yet studied the extent of Wikipedia’s presence on the public Web. In this work, we conduct the first large-scale analysis of how Wikipedia is linked across the Web, something we term it’s “webonization”. We do so by extracting all Wikipedia links from Common Crawl—the largest public dump of HTML content on the Web—and examining three aspects of Wikipedia linking. **RQ1:** What types of Wikipedia articles are linked more vs. less on the Web? **RQ2:** Where are Wikipedia articles likely to be linked (both on the Web in general and within a page)? **RQ3:** Why are Wikipedia articles linked on the Web?

Data and Methodology. The dataset used in this work was extracted from the February 2021 Common Crawl dump¹—the largest public scrape of all HTML pages on the Web created prior to February 2021—through a regex search of <a> tags, the HTML element for hyperlinking. Specifically, the WEB2WIKI dataset contains all the publicly accessible HTML pages that link to Wikimedia projects, as well as the bare webpage–article link pairs. We further use Homepage2Vec to classify webpages into their respective Curlie topics (a clustering of websites) and ORES topics (a clustering of Wikipedia articles) to get coarser representations of articles [1].

Results. *What Wikipedia articles are linked?* We proceed by first tackling the question of what articles are salient on the Web through a comparison of Wikipedia’s Web links to two auxiliary sources. On the one hand, we use Wikipedia in-degree as a proxy of article importance on Wikipedia. In-degree is an effective baseline since it endows an implicit importance score on each article. On the other hand, we define its social importance as the number of times an article is invoked in Reddit comments. Web references, Wikipedia in-links, and Reddit invocations each produce a distribution over the ORES topics. Fig. 1 shows these differences in probabilities for the Web, Reddit, and Wikipedia. We observe that there are some topics which are more likely to be linked on the Web like Computing, Medicine & Health, Philosophy & Religion, and Technology, than other topics like Sports, North America, Northern Europe, Military, and Music which are underrepresented. We also notice some differences between the Web and Reddit, with Reddit being more likely to link to articles belonging to Politics, Philosophy, and Medicine, whereas the Web is relatively more likely to link to articles related to Northern Europe, Music, and Computing. This asymmetry could be indicative of a knowledge gap on Wikipedia between the articles that are important on Wikipedia, and the sites that are important in external discussion. Considering this new article weighting could guide discussions about where to direct attention when curating Wikipedia.

¹<https://commoncrawl.org/2021/03/february-march-2021-crawl-archive-now-available/>

Where Wikipedia articles are linked? Next, we turn to quantifying *where* Wikipedia articles are linked on the Web. Here we measure which webpages are likely to link to a Wikipedia article (using a Curlie trained webpage classification) as well as where within a webpage a link is included (by defining structural HTML rules to determine if a link is in the boilerplate, main content, or a response). Fig. 2 shows the difference in distributions across 14 Curlie topics between websites that link to Wikipedia articles and a random sample of the Web. Since the model provides a separate probability for each class, the probabilities do not necessarily sum to 1. We observe that there exist differences between sections of the Web that link to Wikipedia and the Web as a whole. A larger portion of the Web is made up of Business and Shopping websites, which are less likely to link to Wikipedia articles. On the other hand, News, Science, Society, Arts, Computers, and Reference have a higher proportion of Wikipedia links. We further segment a webpage into three groups: boilerplate, main content, and user responses (e.g. comments). We observe that the vast majority of Wikipedia links occur in the main content (91%), but a non-trivial portion also occurs in the other two segments, with boilerplate and responses representing 7% and 2% of the links, respectively.

Why Wikipedia articles are linked? Finally, we examine why articles are linked using an intuitive taxonomy obtained via iterative coding of Wikipedia links on the Web. To measure the degree to which each type of linking is present, we randomly sampled 500 webpages that link to Wikipedia, and annotated the links based on if they are used for content enrichment or attribution. This manual annotation revealed that content enrichment is by far the largest reason for including links to Wikipedia articles within a webpage (95%), whereas only 5% of links are used for various types of attribution: finding that evidence represents almost 4% of the links, while content sourcing is just over 1%. Generalizing these numbers to all 48.8M English Wikipedia links that are used in the main content of a webpage, we observe that approximately 2.4 million and 46.4 million links are used for attribution and content enrichment, respectively.

Discussion. Research on Wikipedia’s use on the Web is a small but burgeoning area. Studying Wikipedia’s broader connections to the Web has been called upon as a part of the “New Research Directions” by the Wikimedia Foundation in 2015 [2]. More recently, the foundation repeated this call-to-action in its “New Research Roadmap for Identifying Knowledge Gaps” [3]. In their future plan, they argued for studying “knowledge gaps of the broader Web” and the “forms of knowledge Wikimedia needs to acquire in order to fulfill its role on the Web”. By running the first ever large-scale analysis of Wikipedia linking across the Web, this work fills a gap in current research on Wikipedia and the Web, as well as answers fundamental and paramount questions about Wikipedia’s “webonization”. We believe that this second-order approach of analyzing Wikipedia can act as a new lens for studying Wikipedia, and the rich WEB2WIKI dataset presented in this work will be symbiotic with other existing Wikipedia data sources.

- [1] Sylvain Lugeon, Tiziano Piccardi, and Robert West. Homepage2vec: Language-agnostic website embedding and classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 1285–1291, 2022.
- [2] Dario Taraborelli. The sum of all human knowledge in the age of machines: a new research agenda for wikimedia. In *ICWSM-15 Workshop on Wikipedia*, 2015.
- [3] Miriam Redi. A new research roadmap for addressing knowledge gaps, 2022. Available: <https://diff.wikimedia.org/2022/04/21/a-new-research-roadmap-for-addressing-knowledge-gaps/>, Last accessed: 2023-01-14.

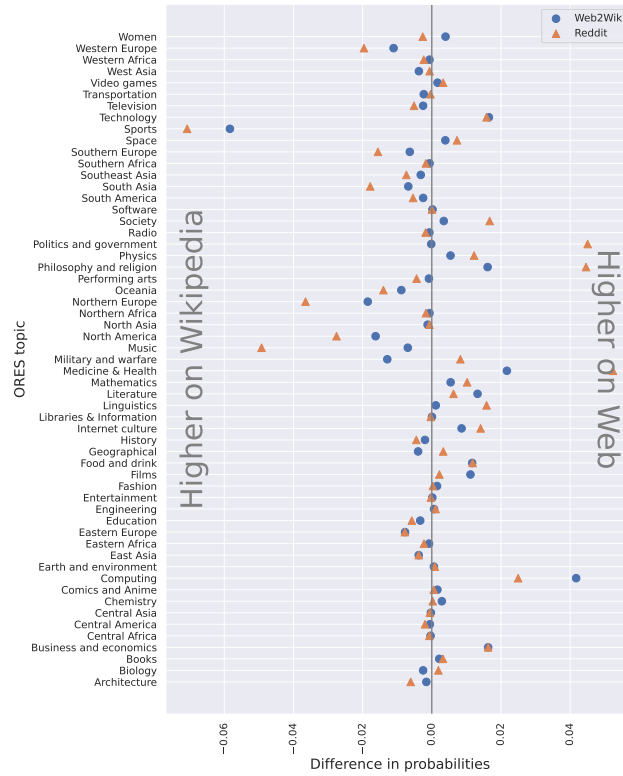


Figure 1: Comparison of proportion of in-links on Wikipedia, Web links, and Reddit invocations of Wikipedia. A higher score means proportionally more in links on the Web than on Wikipedia.

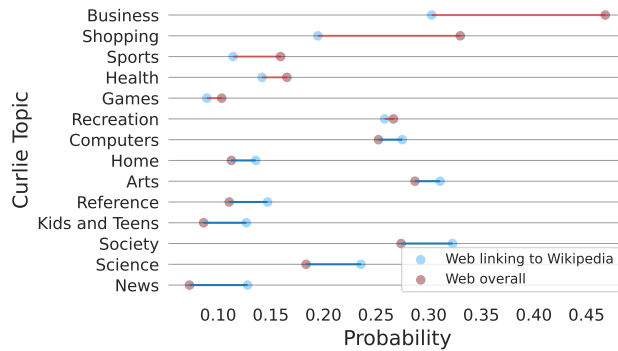


Figure 2: Comparison between websites that link to Wikipedia and a random sample of all websites. For each topic, the difference between the “Web overall” and “Web linking to Wikipedia” probabilities determines its leaning. When the score for “Web linking to Wikipedia” is higher, it implies that the topic tends to link more to Wikipedia, while the reverse implies that Wikipedia articles are underrepresented across that particular topic.