Google2Wiki: A Public Dataset for Mapping the Relevance of Wikipedia across Topics and Time

Bhargav Srinivasa Desikan IPPR
Tiziano Piccardi Stanford University
Martin Gerlach Wikimedia Research
Robert West EPFL

Keywords: Wikipedia, Google, Dataset, Temporal Data Science, Internet Studies

Introduction

Google and Wikipedia are a major part of our web infrastructures and studying their relationship and interdependence is crucial in understanding our online knowledge seeking ecosystems. Wikipedia is the largest online encyclopedia (wik,), and Google holds the highest market share among search engines today. A large fraction of its readers reach Wikipedia from from external search engines (Piccardi et al., 2023), the vast majority of them from Google Search (more than 90%) (Andreescu et al., 2021). In return, Wikipedia articles are believed to improve the quality of the search results (McMahon et al., 2017), (Vincent et al., 2019). However, the nature of this relationship is often in flux - in 2015, with the introduction of the Knowledge Panel for Google Search results, Wikipedia experienced a substantial drop in page views. While the reason for the drop in views has not been verified, various media sources around the world reported this drop as a direct cause of Google's new feature (Orlowski, 2014). Another open question in this relationship is the degree to which Wikipedia articles actually serve the information needs of users online, specifically those using search engines. For example, for information about health, Wikipedia is the most frequently visited resource online (Smith, 2020). However, we lack systematic insights when considering the volume and diversity of content on Wikipedia more generally. Given the changing nature of search on the internet with the advent of LLMs serving as search tools, it becomes increasingly important to track how changes to our internet landscapes effect the largest websites. In the context of Wikipedia and Google, this involves a way to temporally or statically map traffic between the websites. While statistics about the volume of access to Wikipedia articles is publicly available, the raw pageviews are unsuited as they are driven by trends and current events (Miz et al., 2020). The main challenge is to normalize by the overall interest in a specific topic, intrinsically characterized by large fluctuations (Ratkiewicz et al., 2010). However, systematic analysis of click-through rates from search queries to Wikipedia articles is more than challenging. Unfortunately, Google (or other search engines) does not make this data publicly available. Even though more coarse datasets such as Google Trends, which capture the overall volume of specific queries, are publicly available, there are severe technical challenges to, e.g., reconstruct the original signals from normalization and noise perturbation or map queries to Wikipedia articles.

In this work, we develop a novel approach to approximate the click-through rate from Google's search engine to Wikipedia articles that are relevant to a specific query. Our key methodological innovation is to combine, match and model two publicly available resources, Google Trends (i.e., what users on Google are searching for) and Wikipedia clickstream (from where readers reach articles on Wikipedia). Specifically, we make the following contributions:

- We propose a conceptual model that allows us to combine the public data of Google Trends and clickstream data to estimate click-through rates from Google searches to relevant Wikipedia articles. We describe our model and hyperparameters in Figure 1.
- We publish a new multilingual (4 languages) and longitudinal dataset (5 years) of click-through rates from Google for more than 10,000 Wikipedia articles for each edition.
- We publish the code to expand the analysis to further languages in the future.
- We publish early results mapping how Google to Wikipedia rates change over topics and time. We provide an example of a result mapping topics and clickthrough rates in Figure 2.

The data will open the door to systematically studying the relevance of Wikipedia in information seeking online. More generally, it will improve our understanding Wikipedia's role in the larger online ecosystem (Vincent et al., 2018; Piccardi et al., 2021).

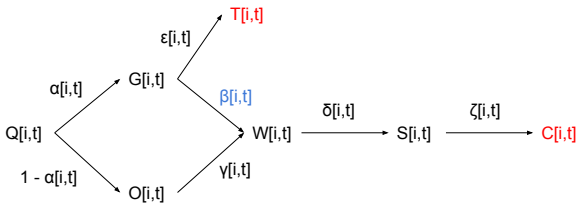

Figure 1: The model in figure 1 describes the relationships between the different values we can collect with public data. Our model reflects the restrictions of Google Trends of comparing all titles across one time period, or comparing one topic across all time periods. It is for this reason that we will consider volumes for the set of all queries made during time t, rather than only query i. In this case, we write $Q[t]$, $G[t]$, $W[t]$, $\beta[t]$, etc. Below are the hyperparameters associated with the model.

Results

Hyperparameters

- i: article index
- t: time (month) index
- Q: number of queries for i across all search engines
- G: number of queries for i on Google
- T: search interest according to GTrends
- O: no. of queries on other search engines
- W: no. of search-originated Wikipedia pageviews of i
- S: no. of search-originated Wikipedia pageviews of i that have SE referrer URL
- C: number of search-originated Wikipedia pageviews of i
- α: Google's market share
- β: fraction of all Google queries that result in a click to Wikipedia
- γ: fraction of all queries that result in a click to Wikipedia, for other search engines
- δ: fraction of search-originated Wikipedia hits with the search engine.
- ζ: fraction of hits that go to Clickstream (either 0 or 1, for a fixed (i,t)).
- ε: $1/N$, where N is the normalizer that Google Trends uses.

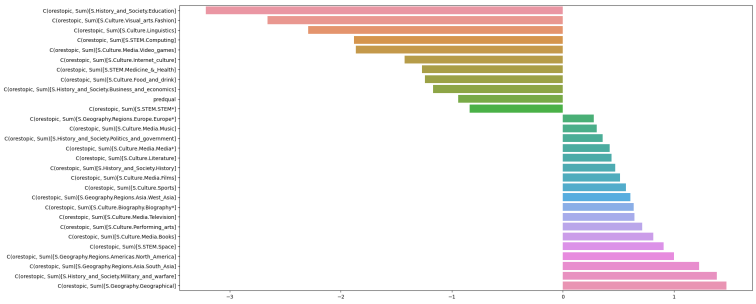

Figure 2: With an ANOVA analysis of google2wikipedia ratios based on groups as ORES topics of titles, we see statistically significant values for differences in the mean beta values for multiple topics; leading the ground for rich interpretation and deeper analysis. For example, given the lack of a popular, authoritative database of cities, towns and countries with associated information, we see pages in the geography topic group rank high in their clickthrough rates - however, for business and economics, it could be the case that more authoritative websites such as the banks or business themselves may be preferred.

References

[Andreescu et al.2021] Dan Andreescu, Kinneret Gordon, Isaac Johnson, and Nicholas Perry. 2021. Searching for wikipedia. `https://techblog.wikimedia.org/2021/06/07/searching-for-wikipedia/`. Accessed: 2023-12-11.

[McMahon et al.2017] Connor McMahon, Isaac Johnson, and Brent Hecht. 2017. The substantial interdependence of wikipedia and google: A case study on the relationship between peer production communities and information technologies. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), May.

[Miz et al.2020] Volodymyr Miz, Joëlle Hanna, Nicolas Aspert, Benjamin Ricaud, and Pierre Vandergheynst. 2020. What is trending on wikipedia? capturing trends and language biases across wikipedia editions. In *Companion Proceedings of the Web Conference 2020*, WWW '20, pages 794–801, New York, NY, USA, April. Association for Computing Machinery.

[Orlowski2014] Andrew Orlowski. 2014. Google stabs wikipedia in the front, January.

[Piccardi et al.2021] Tiziano Piccardi, Miriam Redi, Giovanni Colavizza, and Robert West. 2021. On the value of wikipedia as a gateway to the web. February.

[Piccardi et al.2023] Tiziano Piccardi, Martin Gerlach, Akhil Arora, and Robert West. 2023. A Large-Scale characterization of how readers browse wikipedia. *ACM Trans. Web*, January.

[Ratkiewicz et al.2010] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and

Alessandro Vespignani. 2010. Characterizing and modeling the dynamics of online popularity. *Physical review letters*, 105(15):8–11, October.

[Smith2020] Denise A Smith. 2020. Situating wikipedia as a health information resource in various contexts: A scoping review. *PloS one*, 15(2):e0228786, February.

[Vincent et al.2018] Nicholas Vincent, Isaac Johnson, and Brent Hecht. 2018. Examining wikipedia with a broader lens: Quantifying the value of wikipedia's relationships with other Large-Scale online communities. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, number Paper 566 in CHI '18, pages 1–13, New York, NY, USA, April. Association for Computing Machinery.

[Vincent et al.2019] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the importance of User-Generated content to search engines. *Proceedings of the International AAAI Conference on Web and Social Media*, 13:505–516, July.

[wik] Wikistats - statistics for wikimedia projects. `https://stats.wikimedia.org/`. Accessed: 2023-12-11.