A Deeper Investigation of the Importance of Wikipedia Links to the Success of Search Engines

https://wikiworkshop.org/2020/

Nicholas Vincent and Brent Hecht.

Contact: www.nickmvincent.com | nickvincent@u.northwestern.edu | @nickmvincent
Intelligent Technologies

“Data Labor”

Algorithms & Platforms

Intelligent Technologies
Why Study Data Labor?

● Economic concerns
  ○ Intelligent technologies linked with serious societal harms from inequality
● Recognize and dignify data labor
● Sustainability of peer production

Make people aware of value, to make it possible for them to leverage the value and create change
Why Study Data Labor?

Leverage value of data labor to create a computing paradigm where economic benefits and power are shared much more broadly.

Could range from: paycheck for your data to more recognition/agency for Wikipedia.
So how exactly do Wikipedia and search engines fit into “data labor” research?
Search engines

- ubiquitous
- hugely influential
The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies

Connor McMahon*, Isaac Johnson*, and Brent Hecht2
1*indicates co-First Authors; 2GroupLens Research, University of Minnesota;
2People, Space, and Algorithms (PSA) Computing Group, Northwestern University
mcmah250@umn.edu, isaacj@u.northwestern.edu, bhecht@northwestern.edu

Abstract
While Wikipedia is a subject of great interest in the computing literature, very little work has considered Wikipedia’s important relationships with other information technologies like search engines. In this paper, we report the results of two deception studies whose goal was to better understand the critical relationship between Wikipedia and Google. These broader information technology ecosystem. This ecosystem contains potentially critical relationships that could affect Wikipedia as much as or more than any changes to internal sociotechnical design. For example, in order for a Wikipedia page to be edited, it needs to be visited, and search engines may be a prominent mediator of Wikipedia visitation pat-

26% click through rate

14% click through rate
Wikipedia was a top source of content and appeared in 80-90% of results pages for some categories. 

- not true for every category 
- not always in the top 3 results
takeaway: Wikipedia is one of the most important sources of results for search engines
We had several generalization questions:
● What about search engines other than Google?
● What about mobile results?

Technical challenge:
● How do we handle the ever-changing Search Engine Results Pages ("SERPs") for multiple search engines?
  ○ SERPs are no longer just “10 blue links”
SERPs are not just 10 blue links

Knowledge Box, News Carousel, Twitter Carousel, etc. Presumably very important to search.
Methods
Search Engines, Devices, and Queries

a. What search engines?
   i. Google, Bing, and DuckDuckGo

b. What devices to emulate?
   i. Desktop and Mobile (we also considered the effect of different screen sizes)

c. What queries to make?
   i. Critical and challenging
   ii. Our approach: multiple important categories, drawing on past work
Query selection

“common” queries (100 from search engine optimization company ahrefs.com)

e.g. “facebook”, “youtube”, “amazon”, “gmail”

“trending” queries (282 from Google trends)

e.g. “World Cup”, “thank u, next”, “What is fortnite”

“medical” queries (50 from prior research that shared Bing data)

e.g. “how to lose weight”, “indigestion”

See:

Data collection

Our approach: use `puppeteer` (Node.js) run “headless” Chrome

- We forked NikolaiT’s `se-scraper` library: https://github.com/NikolaiT/se-scraper
- Our version focuses on recording and analyzing link coordinates with the space of a single SERP
- I’ll repost code links on the final slide
One approach for SERP scraping:

Researcher looks at SERP HTML

Write CSS rules to parse HTML page into a ranked list

“find all elements with class of searchResults_abc123”

1. wikiworkshop.org
2. twitter.com/wikiworkshop
How should we turn this into a ranked list?
Spatial analysis: getting link coordinates

Get all the links (“a” elements) in a page:

```javascript
await this.page.$$eval('a', getPos);
```

Calculate their position (x, y) with JavaScript:

```javascript
Element.getBoundingClientRect()
```
Spatial incidence rate definition
Incidence rates

- how often is Wikipedia showing up in SERPs?
  - If we collect 3 SERPs, and Wikipedia appears twice, incidence rate = 2 / 3
- how often is Wikipedia showing up in prominent parts of SERPs?
  - If Wikipedia appear “above-the-fold” in only one of our 3 SERPs, above-the-fold incidence rate = 1 / 3
Data validation - a tough task

SERP data changes constantly - remember this?

Google is walking back changes to its search design that blurred the lines between ads and regular results after user backlash

Tyler Sonnemaker  Jan 24, 2020, 1:31 PM
Data validation - visual inspection

Basic approach:

- save screenshots of SERPs
- visualize the analysis-ready data (i.e. the JSON files for quantitative analysis)
- make sure they seem to match up!
Results
Full page incidence rates
Above-the-fold incidence rates

y-axis = Above-the-fold incidence | Device = desktop

y-axis = Above-the-fold incidence | Device = mobile

Search Engine
- google
- bing
- duckduckgo

Incidence rate

Query Category:
- common
- trending
- medical
Left-hand and right-hand
Summary of findings

● Using the easy-to-understand (but limited) measure of incidence rates, Wikipedia’s importance to the success of search engines extends beyond Google and desktop-formatted search results
● Queries and devices matter:
  ○ Wikipedia appears above the fold more often on desktop devices than mobile devices
  ○ Knowledge panel elements are a key source of Wikipedia content, but not the only sources
Data from the Public Fuels Intelligent Technologies

- Are Wikipedia editors some of the most important employees of search engines?

You **cannot** pay people to edit Wikipedia.


Positive: effective peer production = effective search results?

Negative biases in coverage / quality = impact on search results?

Raises the stakes of Wikipedia-focused research and Wikipedia findings
Limitations

● Audit study!
  ○ Small scale relative to actual query datasets

● Still US / en.wikipedia only
  ○ Wikipedia has geographic / language differences

● Queries matter immensely.
  ○ Incidence rate can lose meaning very quickly, e.g. if we append “wikipedia” to each query
  ○ Interpret accordingly
Many thanks to:

co-author Brent Hecht, thoughtful reviewers and colleagues

Open Software

- se-scrapers / puppeteer
- numpy / pandas / seaborn / scipy ecosystem
Community Data Science Collective’s COVID-19 Digital Observatory

-has SERP data for COVID-related keywords, using a newer version of this software developed after this study
-other data of interest to this group (Wikipedia editing, reddit)
Questions?
@nickmvincent on Twitter or nickvincent@u.northwestern.edu

Links:
paper: https://www.nickmvincent.com - under “Pre-prints”
se-scraper: https://github.com/NikolaiT/se-scraper
our se-scraper fork: https://github.com/nickmvincent/se-scraper
updated scraping repo: https://github.com/nickmvincent/LinkCoordMin
COVID-19 Digital Observatory:
https://wiki.communitydata.science/COVID-19_Digital_Observatory