Visual Positions of Links and Clicks on Wikipedia

Dimitar Dimitrov
GESIS – Leibniz Institute for the Social Sciences
dimitar.dimitrov@gesis.org

Philipp Singer
GESIS – Leibniz Institute for the Social Sciences
philipp.singer@gesis.org

Florian Lemmerich
GESIS – Leibniz Institute for the Social Sciences
University of Koblenz-Landau
florian.lemmerich@gesis.org

Markus Strohmaier
GESIS – Leibniz Institute for the Social Sciences
University of Koblenz-Landau
markus.strohmaier@gesis.org

ABSTRACT

In this work, we study the visual position of links and their clicks on Wikipedia, particularly where links are visually located, at which screen positions users click on links, and which areas on the screen exhibit more or less clicks per links. For that purpose, we introduce a novel dataset containing the on-screen coordinate position for all links between pages in the English Wikipedia and additionally resort to navigation logs of Wikipedia users. Using this data, we can observe a preference of certain link and click locations on Wikipedia including first evidence of positional click bias. For example, our results suggest that users have a tendency to prefer to click on the left side of the screen which exceeds what one would expect from the presence of links on pages. We believe that presented data and research can be useful for optimizing the process of link creation and link consumption on Wikipedia and other Web platforms.

Keywords: Human Click Behavior; Navigation; Wikipedia

1. INTRODUCTION

In this work, we study the visual position of links and clicks on Wikipedia; we investigate *where links are positioned*, *where users click on links* and *which regions expose more (or less) clicks per links*. Our main contributions are twofold: (i) We introduce a novel dataset capturing the visual position of all links between articles of the English Wikipedia (including also templates, infoboxes/sidebars, and navboxes) based on their fine-grained screen coordinates—the basic visual structure of a Wikipedia page is shown in Figure 1. (ii) We present first empirical insights into the position of links and clicks on a screen providing first evidence for the existence of positional click bias on Wikipedia. Our findings suggest that links on Wikipedia pages are not created and consumed equally, e.g., the preference of users to click on the left side of the screen exceeds what one would expect from the presence of links on pages. Thus, the results from any method tapping into the Wikipedia network topology might be influenced by the position of links on a page.

While the influence of link position on click rates has been studied before in the context of search engine log analysis [1] and mouse cursor behavior on search engine results pages [2], this has not been analyzed for Wikipedia yet. Our work shows that the preference of links based on their displayed screen position is also prevalent on information networks such as Wikipedia.

WWW'16 Companion, April 11–15, 2016, Montréal, Québec, Canada.
ACM 978-1-4503-4144-8/16/04.
http://dx.doi.org/10.1145/2872518.2889388.

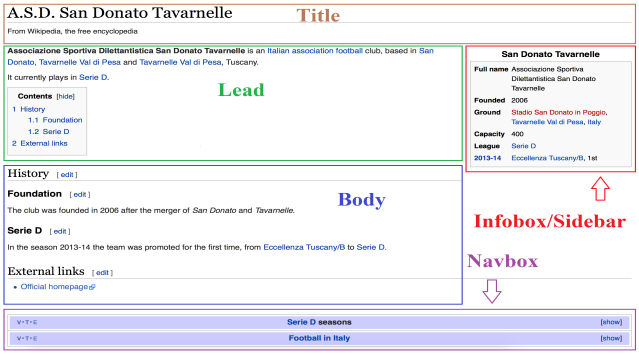


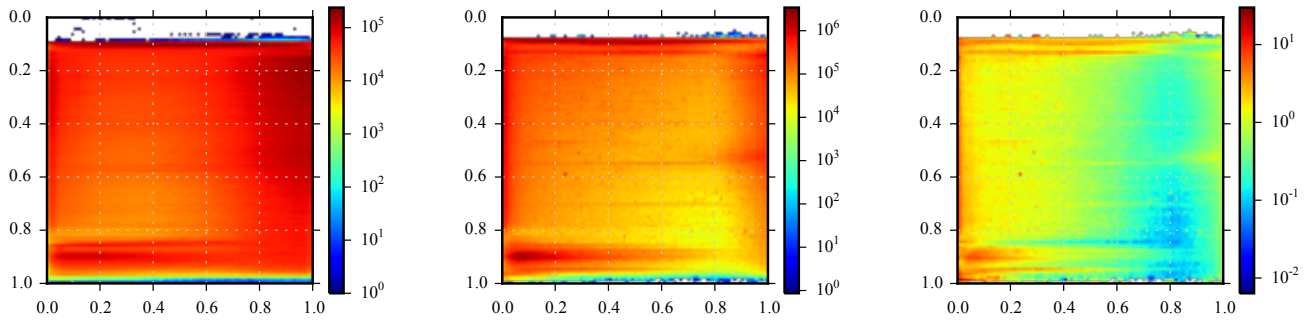
Figure 1: ***Visual structure of a typical Wikipedia page.***

2. DATASETS AND METHODS

Wikipedia link dataset. Our contributed dataset includes information about all Wikipedia articles that are part of the English Wikipedia XML dump from March 4th, 2015. To avoid rendering issues with the wikitext contained in the XML dumps, we additionally obtained the full HTML versions of all articles via Wikipedia's new REST API[1]. We then extracted all internal link records on a Wikipedia page by processing the HTML files; for resolving redirects, we utilized the XML dump. We rendered each page using the PyQt4 framework[2] which allowed us to derive the precise screen coordinates of each link, i.e., the coordinate of the upper left corner of a link as displayed on-screen. For rendering, we decided to use the commonly used high definition resolution of 1920×1080 pixels as we focus on desktop users in this study, see the description of the clickstream dataset below. The final link-position dataset contains $4,805,500$ articles connected by overall $429,917,702$ links and $340,126,041$ distinct links (some links occur multiple times on a page) and provides a screen coordinate for each link.

Wikipedia clickstream dataset. We additionally use a recently published Wikipedia clickstream dataset from February 2015 [3]. This dataset contains about 22 million *(referrer, resource)* pairs and their respective request count (≥ 10) extracted from the request logs of the main namespace of the desktop version of the English Wikipedia. The dataset features overall 2.3 billion requests. Requests that were made at too high of a rate were discarded as well as requests identified to be incited by bots or spiders, see [3] for more details about the applied data cleaning. The referrers can be categorized in internal and external traffic; in this work, we only

[1] `https://en.wikipedia.org/api/rest_v1/?doc`
[2] `http://pyqt.sourceforge.net/Docs/PyQt4/`



(a) *Links heatmap* (b) *Clicks heatmap* (c) *Clicks/Links heatmap*

Figure 2: *Heatmaps*. (a) shows links positions, (b) clicks positions, and (c) clicks/links. The links heatmap (a) indicates high link density in the lead, the infobox/sidebar, and the navbox regions. The clicks heatmap (b) shows the regions with high click frequency—the lead, the infobox/sidebar, navbox and left body. The clicks/links heatmap (c) highlights the preference of users clicking on the left side of the screen, exceeding expectations implied by the presence of links. All heatmaps are logarithmically scaled.

focus on request pairs stemming from internal Wikipedia traffic, i.e., referring page and requested resource are both Wikipedia pages from the main namespace. To map the clickstream dataset to our link position dataset, we only consider request pairs that had both the referrer and the resource present in the link dataset. This leads to overall $13,622,339$ distinct pairs featuring $1,435,738,382$ user transitions between articles.

Method. For empirical insights into visual link and click positions, we adopt heatmaps that are calculated by dividing the screen into 100×100 equally sized bins and then counting the number of times (i) a link exists and (ii) a link is clicked in the respective bin. In order to normalize screen width and height, we divide the x coordinate of a link by the screen width (1920) and the y coordinate of a link by the length of the respective page. We ignore links from the HTML that are not visible due to css styling. When multiple links with the same target exist on a page, we divide the actual click count by the number of link occurrences on the page as we do not know which of the links people actually clicked. For (iii) studying which regions produce more or less clicks per link, we re-use the heatmaps for clicks and links and perform an element-wise division of the corresponding bin counts. The counts in all bins of the three heatmap are logarithmically scaled for the color mapping.

3. RESULTS AND DISCUSSION

Results. Figure 2 shows three different heatmaps based on the data and methodology described in Section 2. Figure 2(a) shows the position of links on the screen. When comparing the results to the visual structure of a typical Wikipedia page as shown in Figure 1, we can identify four main concentrations of links on the screen: (i) the lead section on the top, (ii) the section on the right hand side containing infoboxes and sidebars, and (iii) the bottom area mostly containing navboxes (specifically, the left part). In comparison, the main body shows lower density with exception of (iv) the outer left part which might be explained by the multiple presence of lists with many links in Wikipedia articles. Figure 2(b) shows the location of clicks on the screen. People seem to prefer to click on those links that are located (i) at the top of the screen in the lead section, (ii) on the right sidebar (focus on infoboxes), (iii) in the bottom navboxes (focus on left side), as well as (iv) the left side of the body of a page. Not surprisingly, these patterns are overall similar to those observed in Figure 2(a) since users can only click on links that exist. Yet, some differences can be detected such as a general disfavor of regions located at the right hand side of the screen. To further investigate these differences, 2(c) displays the number of clicks per links in a region. Here, hot colors indicate regions with high numbers of clicks per link, cool colors the opposite. The results confirm our intuition that people prefer to click on links in the left side of the screen, and this preference exceeds what one would expect from the link count in that area. By contrast, the links in the sidebar on the right hand side (infoboxes) are less often followed as the pure number of links there would suggest.

Discussion. In this work, we have presented a novel dataset that captures the visual position of links in Wikipedia. First empirical results suggest that links in Wikipedia pages are not produced and consumed equally. Next, we shortly want to highlight some limitations, ideas for future work and implications of this work.

Limitations. (i) The utilized click data only provides an approximation of clicks. Future work could also contrast these studies by analyzing more fine-grained data such as eye-tracking or mouse movement studies. (ii) Due to this approximated data, we also have the issue of not always being able to unambiguously assign a click to a specific link position as Wikipedia pages may contain multiple links with the same target. We have addressed this by assigning each possible position an equal amount of attention. However, more refined approaches are warranted in future work. (iii) We chose a specific screen resolution in this work for deriving and studying the visual link positions. Future work should extend this research to various resolutions for more detailed insights. Our framework can work with arbitrary screen resolutions. (iv) Our empirical insights give an aggregated view without distinguishing between individual pages. In future work, we plan a more detailed study aiming at the impact of link positions on an individual page level.

Implications. The data and empirical insights presented in this work should encourage future research in the direction of studying the visual presence of links and their usage. We believe that this can be useful for optimizing the process of link creation in Wikipedia—e.g., by adapting the link creation guidelines—and other Web platforms, but also for creating or improving models for human navigation in information networks.

References

[1] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey. An experimental comparison of click position-bias models. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*. ACM, 2008.

[2] J. Huang, R. W. White, and S. Dumais. No clicks, no problem: Using cursor movements to understand and improve search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011.

[3] E. Wulczyn and D. Taraborelli. Wikipedia clickstream. figshare. `doi:10.6084/m9.figshare.1305770`. Accessed: 2015-12-13.