Quo Vadis? On the Effects of Wikipedia's Policies on Navigation

Daniel Lamprecht
Knowledge Technologies Institute
Graz University of Technology
daniel.lamprecht@tugraz.at

Denis Helic
Knowledge Technologies Institute
Graz University of Technology
dhelic@tugraz.at

Markus Strohmaier
GESIS and University of Koblenz-Landau
markus.strohmaier@gesis.org

Abstract

We propose to study the influences of Wikipedia's policies on navigation in Wikipedia and describe our methods to study navigational biases, assess the guidelines provided by the Manual of Style, and investigate the neutrality of navigation.

Introduction

When browsing the Web, users typically have certain expectations and are influenced by cognitive biases. One such example is position bias: Humans are known to dedicate more of their attention towards the top of a page or a list (Payne 1951; Salganik, Dodds, and Watts 2006; Lerman and Hogg 2014). Users generally scan Web pages in an F-shaped pattern (Nielsen 2006), dedicating more time to areas where they expect to find the most important elements such as a navigation bar or the introduction.

On the English Wikipedia, one of the most-visited Websites worldwide, articles tend to have a rather rigid structure. For example, the first phrase usually puts the article title in context to more general concepts, and category links are generally located at the bottom of the page. Because of this fixed structure, users likely have certain expectations about where to find links or specific pieces of information. In the first part of our analysis, we study expectations and biases present in Wikipedia navigation by comparing a range of biased navigation models to a Wikipedia clickstream.

The policies and guidelines Wikipedia editors follow to structure articles are developed by the community and collected in the *Manual of Style* (Wikipedia 2015a). Many of these policies also affect navigation on the Wikipedia network. As an example, categories, lists and navigation templates are all navigational aids described in the Manual of Style. All three elements are purposely redundant—all can be used to navigate to related articles, and they are intended for different user preferences. In the second part of this work, we analyze the influence of these policies and guidelines on navigation in Wikipedia by evaluating the click frequencies to structural elements intended for navigation.

One of the most fundamental principles of Wikipedia is its neutral point of view (Wikipedia 2015c), which the community strives to enforce on articles. While great care has been dedicated to the neutrality of Wikipedia's textual content, neutrality in terms of the Wikipedia link network has received comparatively little attention. In the third part of this work, we analyze Wikipedia's neutral point of view from a navigational perspective and investigate to what extent this principle holds true for notions of navigation such as reachability or centrality in the Wikipedia link network. We assume that for concepts of equal importance and familiarity, such as the two major candidates of a U.S. presidential election, Wikipedia would expect both to be reachable equally well. To evaluate this hypothesis, we first analyze the paths leading to both of these articles and then take into account user expectations and biases.

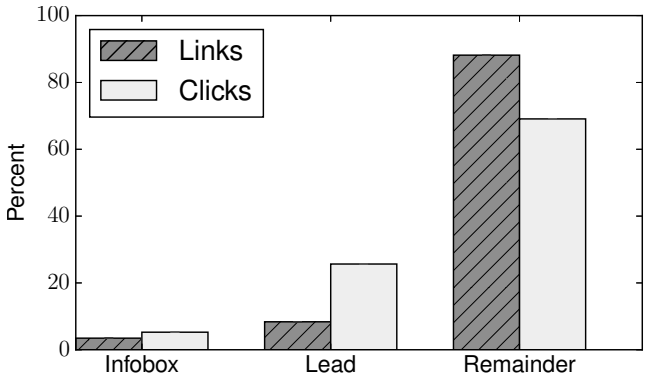

Figure 1: **Fractions of links and Clicks.** The figure shows the links as a fraction of the total number of links present and the clicks as a fraction of total clicks. In case of multiple occurrences of a link, we assumed clicks were equally distributed. This figure uses data from roughly 4600 articles from the English Wikipedia. The figure shows that links located near the the top of pages were clicked more frequently.

Contributions: We conduct a broad evaluation of the effects of Wikipedia's policies on navigation, investigate the usage of a range of navigational aids in Wikipedia articles and study the question of neutrality in navigation and reachability. Our results constitute a first step towards an evaluation of Wikipedia's policies, specifically in terms of Wikipedia's link network.

Materials and Methods

We use data from the **February 2015 English Wikipedia Clickstream** (Wulczyn and Taraborelli 2015), which consists of all clicks from and to articles in the English Wikipedia in February 2015. We only use data for clicks leading to and from Wikipedia articles and drop all clicks from and to external Web sites.

Comparison of navigational influences

To investigate influences on navigation, we use Markov chains based on several potential navigational influences. Specifically, we construct first-order Markov chains, which have been found to be a fitting model for Web navigation (Singer et al. 2014). We first construct a Markov chain for the English Wikipedia and compute transition probabilities from the clickstream data. Next, we model a range of navigational influences as Markov chains as well and evaluate how well they are able to explain the clickstream data. We investigate the following influences:

- Biased towards selecting a link in one of the areas of interest (infobox, lead, ...)
- Bias to link target generality (as measured by indegree)
- Bias to popularity (as measured for example, by frequency of occurrence in search engine queries)
- Random (serving as a baseline)

To compare these to the clickstream, we compute the stationary distributions of the Markov chains and compute their correlation. We then evaluate a combination of these influences to best fit the clickstream data.

Assessment of navigational aids

As a first step towards the evaluation of navigational aids on Wikipedia, we empirically analyze the distribution of clicks to the infobox and the lead section. We use data from roughly 4600 articles (matching those of the 2007 Wikipedia for Schools selection), which we obtained from the English Wikipedia in March 2015. Figure 1 shows that generally speaking, the vast majority of links (over 90%) are located outside the infobox and lead section. However, weighting links by their click frequencies shows that 26% of clicks occur inside the lead section.

By Wikipedia policies, *a link should appear only once in an article [...] but may be repeated in infoboxes, tables [...] and at the first occurence after the lead* (Wikipedia 2015b). We find that this leads to repeated occurrences for around 25% of links in our data. As the clickstream dataset does not include information on the exact position of clicks, we distribute click frequencies uniformly to all occurrences of a link. Since Web users have been found to dedicate more attention to the top of pages (Nielsen 2006; Lerman and Hogg 2014), this approach is likeley overly conservative. When we assume clicks always occurred on the first occurrence of a link and repeat the analysis, the fraction of links clicked in the lead section increases to 43%. It appears likely that true fraction lies somewhere in between 26% and 43%.

This first result shows that Wikipedia users dedicate a large share of their attention to the lead section. Policies affection the lead section of articles therefore have a substantially larger effect on issues of navigation.

We intend to extend this analysis to further areas of interest, such as the first 1000 words in an article, lists, categories, and navigation templates.

Neutral Navigation

Writing articles from a neutral point of view is one of Wikipedia's core principles, and articles are expected to present all significant viewpoints in proportion to representation in reliable sources on the subject (Wikipedia 2015c). While the Manual of Style dedicates much attention to textual neutrality, the subject of neutrality in links is only briefly touched upon. In this analysis, we investigate neutrality in terms of navigation. We start with comparing the reachability of several pairs of high-profile articles such as the two major candidates for the U.S. presidential election. We distinguish two cases:

- For articles for which we assume equal importance in February 2015 (such as the two candidates for the 2015 Chicago mayoral election), we count the number of visit as detailed by the February 2015 Wikipedia Clickstream dataset.
- For articles with presumably equal importance before February 2015 (where we do not have detailed clickstream data), we count the number of inlinks, weight by overall visit counts of the referring pages and compute PageRank centralities.

We intend to perform this analysis to a range of comparable topics, such as competing corporations (e.g., Airbus and Boing) or brands (e.g., Nike and Adidas), where we would assume public interest and familiarity to be approximately equal. Our investigation will also take into account the importance of links, as detailed by our analysis of biases and the Wikipedia clickstream dataset.

Proposed Objectives

We aim to come up with both an assessment of the current state of navigability on Wikipedia as well as a set of suggestions for future improvements. For example, a simple remedy could be the removal or addition of links, guided by methods similar to (West, Paranjape, and Leskovec 2015). More subtle changes could be made by adapting the perceived importance of links based on our study of navigation biases exhibited by Wikipedia users—for example, by moving a link towards the top of the page, or by adding an article to an additional category, thus increasing its visibility to users and making it more likely to be clicked on.

Acknowledgments

This research was supported in part by a grant from the Austrian Science Fund (FWF) [P24866].

References

Lerman, K., and Hogg, T. 2014. Leveraging position bias to improve peer recommendation. *PLOS ONE* 9(6):e98914.

Nielsen, J. 2006. F-shaped pattern for reading web content. http://www.nngroup.com/articles/f-shaped-pattern-reading-web-content. [Online; accessed 31-March-2015].

Payne, S. L. 1951. *The Art of Asking Questions*. Princeton University Press.

Salganik, M. J.; Dodds, P. S.; and Watts, D. J. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.

Singer, P.; Helic, D.; Taraghi, B.; and Strohmaier, M. 2014. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PLOS ONE* 9(7):e102070.

West, R.; Paranjape, A.; and Leskovec, J. 2015. Mining missing hyperlinks from human navigation traces: A case study of wikipedia. In *Proceedings of the 24th international conference on World Wide Web*. ACM. forthcoming.

Wikipedia. 2015a. Wikipedia: Manual of style. https://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style&oldid=654332151. [Online; accessed 31-March-2015].

Wikipedia. 2015b. Wikipedia: Manual of style/linking. https://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style/Linking&oldid=650644240. [Online; accessed 31-March-2015].

Wikipedia. 2015c. Wikipedia: Neutral point of view. https://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=654031907. [Online; accessed 31-March-2015].

Wulczyn, E., and Taraborelli, D. 2015. Wikipedia Clickstream. http://dx.doi.org/10.6084/m9.figshare.1305770.