

A Data-Driven Sketch of Wikipedia Editors*

Robert West
Stanford University
west@cs.stanford.edu

Ingmar Weber
Yahoo! Research Barcelona
ingmar@yahoo-inc.com

Carlos Castillo
Yahoo! Research Barcelona
chato@acm.org

ABSTRACT

Who edits Wikipedia? We attempt to shed light on this question by using aggregated log data from Yahoo!'s browser toolbar in order to analyze Wikipedians' editing behavior in the context of their online lives beyond Wikipedia. We broadly characterize editors by investigating how their online behavior differs from that of other users; e.g., we find that Wikipedia editors search more, read more news, play more games, and, perhaps surprisingly, are more immersed in pop culture. Then we inspect how editors' general interests relate to the articles to which they contribute; e.g., we confirm the intuition that editors show more expertise in their active domains than average users. Our results are relevant as they illuminate novel aspects of what has become many Web users' prevalent source of information and can help in recruiting new editors.

Categories and Subject Descriptors: H.1.2 [Models and Principles]: User/Machine Systems—*Human factors*.

General Terms: Experimentation, Human Factors, Measurement.

Keywords: Wikipedia, editors, Web usage, expertise.

1. INTRODUCTION

Despite Wikipedia's obvious importance, little is known about its editors, with the primary sources being user surveys and analyses of editors' activity within Wikipedia. For instance, according to the Wikimedia Foundation's 2011 yearly editor survey [1], answered by more than 5K editors, they are well educated, with 61% having a college degree and 72% reading Wikipedia in more than one language; the median age is 28 years. In this paper, we introduce a new source of information: traces of browsing behavior. We use data obtained by Yahoo! Toolbar and look for URLs corresponding to Wikipedia edits. This way we can obtain insights into the browsing behavior of Wikipedia editors both in general and during the time period immediately preceding an edit event.

We find that, on broad average, Wikipedia editors seem, on the

one hand, more sophisticated than usual Web users; on the other hand, they are also deeply immersed in pop culture. We also confirm the intuition that editors are typically experts in their domains.

Characterizing Wikipedia editors may be useful from a practical perspective. In order to target promising readers for converting them into editors, it can help a lot to know what a typical editor is like. In this respect, organizations such as the Wikimedia Foundation can directly profit from the results of our research.

2. DATA-SET DESCRIPTION

Users of Yahoo! Toolbar have the option to allow Yahoo! to collect information about the websites they visit, including the user's unique toolbar id, the timestamp, the URL of the page and its referer, and locale information. We consider a large toolbar sample and identify Wikipedia edits in the data by their characteristic URL pattern. In order to eliminate false positives, and to collect additional information about the edit (such as its size and the user's Wikipedia name), we use the timestamps to look up all candidates in the Wikipedia edit logs and keep only those for which we find a match.

The set of all users is divided into 3 groups: editors of the English Wikipedia (1.9K, or .089% of all users), readers-only of the English Wikipedia (58%), and those that do not read any language version of Wikipedia (41%). We subsample 5K readers-only and 10K non-readers (from primarily English-speaking locales), in order to have roughly equal numbers of pageviews.

Our fundamental unit of analysis is that of an *editor–article pair* (EAP), which collapses all edits a given user made to a given article. We define the edit size of an EAP as the maximum edit size over all its constituent edits, measured as the number of bytes in the article after the edit, minus before the edit. All in all, we have around 13K atomic edit events, grouped into 5.3K EAPs.

We address the concern of population bias by comparing the distribution over topics of edited articles in our data set with that of a random sample of recent edits on live Wikipedia, finding a high Pearson correlation coefficient of $r = 0.88$.

3. WHO ARE THE WIKIPEDIANS?

In our data, editors have on average 3 times as many pageviews as readers-only, and 9 times as many as non-readers. In a more fine-grained analysis, we look at how the 3 groups differ in terms of the Web domains they frequent. We represent each user by a *relative domain frequency vector*, which counts for each of the 10K most popular domains (as of Sept. 2010, according to Alexa) what fraction of all their pageviews they spent on it. To have interpretable results, these domains were then grouped into categories. In some cases, this grouping was done by simply taking the top-level domain (e.g., .edu) or by searching URLs for a particular pattern. But in most cases, we used all domains listed in the Yahoo! Directory for the respective category (e.g., ENTERTAINMENT/GAMES).

*Part of this work was done while the first author was visiting Yahoo! Research Barcelona under the Yahoo! internship program. This research has received funding from the Spanish Ministry of Science and Innovation through project CEN-20101037 CENIT Social Media, the EC's Seventh Framework Programme (FP7/2007-2013) under grant agreement 288024 (LiMo-SiNe project), and the Torres Quevedo program.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WWW'12, April 16–20, 2012, Lyon, France.
Copyright 2012 ACM ...\$10.00.

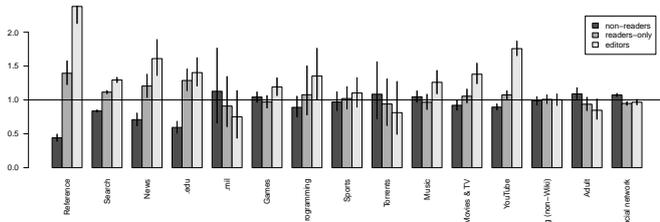


Figure 1: Domain category frequencies (with 95% confidence intervals) for the 3 user groups, each macro-averaged over users. The line at 1.0 corresponds to the overall average for general Web users. All fractions are normalized by this global average, so we can plot everything in one figure.

Fig. 1 summarizes the differences between the 3 groups with respect to the most interesting domain classes. We observe that Wikipedia editors are smart (more news, educational domains, lookups on reference sites, and searches) but fun (more YouTube, music, games, and TV). They also have a lower fraction of pageviews on adult content and social networking sites. Contrarily to Wikimedia’s survey [1, p. 3], we find that editors have a higher fraction of pageviews on game sites than the average. Also, the survey states that a typical editor is ‘computer savvy but not necessarily a programmer’; indeed, we find that editors have significantly more than average pageviews on programming websites.

The fact that editors seem to be more immersed in pop culture is in tune with the topic distribution of the edited articles: around 40% of them have the top-level category ENTERTAINMENT in Yahoo! Directory. It is therefore an interesting question if the entertainment bias is characteristic of all editors or just of those that edit the many entertainment articles. We find that, while entertainment editors have a significantly higher fraction of entertainment pageviews than non-entertainment editors, the latter spend in turn more pageviews on entertainment sites than non-editors (e.g., YouTube: 9.2% vs. 7.7% vs. 3.8%; MOVIES & TV: 2.8% vs. 1.9% vs. 1.4%).

While the above analysis compares editors to non-editors, we now want to see how homogeneous the group of editors is. To this end, we perform principal component analysis (PCA) on users’ relative domain frequency vectors. The first principal component—explaining on its own 47% of the total variance—has values close to 0 in all dimensions except `facebook.com`, with a value of around 11. This lets us conclude that the main line of divide is Facebook. Users from the cluster with more Facebook activity produce more entertainment edits (47% vs. 40% of EAPs), whereas the other cluster produces more edits in SCIENCE (6.3% vs. 4.2%) and NEWS & MEDIA (4.7% vs. 2.5%). Users from the less-Facebook cluster are more involved in Wikipedia, with larger edits (mean/median size 200/45 vs. 123/33), a higher chance of being logged in to Wikipedia (26% vs. 16%), and more edits per user (3.0 vs. 1.8).

4. ARE WIKIPEDIA EDITORS EXPERTS?

Expertise. We now concentrate on the group of editors and investigate if they are experts in the areas in which they make edits. First, we define an editor e ’s *interest* in a Wikipedia article a as the mean similarity between e ’s search queries and a (the definition of similarity between an article and a query is rather technical; suffice it to say it is derived from their Yahoo! Directory categories). Then, we define e ’s *expertise* in a as the ratio of e ’s interest in a to the average editor’s interest in a . Intuitively, someone is an expert in a topic if their interest is significantly above average. Macro-averaged expertise (i.e., first aggregating all EAPs of the same user) is 1.5. That is, an edit is on average one and a half times as related to the given editor’s personal query history as to a random editor’s, indicating that editors know more about the topics they edit than

average Internet users. However, expertise is not equally present in all categories; e.g., it is 4.2 for SPORTS and only 1.5 for ARTS.

Co-expertise. We also investigate the question of *co-expertise*, i.e., whether editors tend to be experts in domains they do not edit. In summary, we state that, among all editors, expertise in the ENTERTAINMENT domain resides mostly in the group of editors of that category, and editors of ENTERTAINMENT typically have no other areas of strong expertise. Hence, the simplistic image of all Wikipedia editors being entertainment-loving has to be faceted: rather, the overall focus on entertainment may be attributed to a group of entertainment-only specialists that contribute many edits. On the contrary, editors of SCIENCE and BUSINESS are more versatile: they show expertise in several areas beyond what they edit.

Correlates of expertise. Some quantities correlate with the expertise of edits; e.g., long edits (notably the very long ones) come from editors with more expertise—a good sign: small edits are often minor corrections such as typo fixes, while the large ones are the real content contributions, which we would hope to come from real experts. Also, editors that use a Wikipedia account show more expertise, implying that the more involved users are better experts. This is confirmed by further findings: among those logged in, those with a barnstar (a Wikipedia-internal award) have significantly more expertise. Finally, it seems that the ‘newbies’ have less expertise: editors who have ever only made one edit and those who have been registered for at most one day (i.e., they probably registered to make the edit in our toolbar logs) have the lowest expertise.

5. WHAT HAPPENS BEFORE AN EDIT?

Additionally, we analyze expertise from a temporal perspective and find that editors show significantly more interest (as expressed in their search queries) in the topic of the edited article just before an edit than usually: when computed based only on the queries during the 30 minutes before the edit, expertise is 4 times as high as when computed based on a random sample of the same editor’s queries. One obvious reason for this is that the query immediately before the edit is often a navigational query issued to find the edited page. But even when we exclude this last query, expertise is still 3 times as high as usually.

Note, however, that the edited article is not always reached via search engines: only 47% of all referral chains culminating in edit events originate on search engines, while 49% start directly on Wikipedia, with the user typing the URL directly or having bookmarked it (4% originate on other pages). Surprisingly, 30% of the referral chains starting on a search engine proceed to the edit immediately, without further browsing.

6. CONCLUSIONS

This paper attempts to shed light on a question that is of importance in order to understand the phenomenon of Wikipedia: Who are the people contributing to it? In particular, we try to draw the portrait of Wikipedia editors in a data-driven fashion, and approach the question from many different angles, characterizing their typical Web usage patterns and their levels of expertise. Knowing typical characteristics of the people active on Wikipedia is useful in designing strategies for retaining editors and turning readers into first-time editors. In doing so, it is important to keep in mind that, while editors as a whole seem to expose certain traits, such as being more sophisticated than average Web users, there is no quintessential Wikipedia editor. Therefore, Wikipedia—as a collaborative system—should continue striving to accommodate diversity.

7. REFERENCES

- [1] Wikimedia Foundation. Editor survey, April 2011. [Online; accessed September 13, 2011].