Approaches to Understanding Indigenous Content Production on Wikipedia

Manasvini Sethuraman
msethuraman3@gatech.edu
Georgia Institute of Technology

Rebecca E. Grinter
beki@cc.gatech.edu
Georgia Institute of Technology

Ellen Zegura
ewz@cc.gatech.edu
Georgia Institute of Technology

ABSTRACT

With over 6 million articles and 38 million editors, English Wikipedia is one of the largest peer produced and edited encyclopedias on the Internet. Some attribute this success to the peer production process which allows many to contribute. However, not all articles are created equally, or evolve similarly, in terms of quality and amount of attention they receive from the editing community. We examine pages with geo-tagged content in English Wikipedia in four categories, places with Indigenous majorities (of any size), Rural places, Urban Clusters, and Urban areas. We find significant differences in quality and editor attention for articles about places with Native American majorities, as compared to other places.

CCS CONCEPTS

• **Human-centered computing** → **Collaborative content creation**; *Empirical studies in collaborative and social computing*; Computer supported cooperative work.

ACM Reference Format:

Manasvini Sethuraman, Rebecca E. Grinter, and Ellen Zegura. 2020. Approaches to Understanding Indigenous Content Production on Wikipedia. In *ACM SIGCAS Conference on Computing and Sustainable Societies (COMPASS) (COMPASS '20)*, June 15–17, 2020, , Ecuador. ACM, New York, NY, USA, 2 pages. <https://doi.org/10.1145/3378393.3402249>

1 INTRODUCTION

Wikipedia is one of the most well-known success stories of peer produced and reviewed content on the Internet. Research has shown that Wikipedia's success in producing not only a significant quantity of information, but has also shown that by some metrics, the quality of that content is also good [2, 6]. However, questions still remain about the imbalances in the amount of content produced and its quality, and the relationship of those differences to external phenomena such as urban/rural digital divide [5]. Wanting to address the question of how rurality and race interact in the digital divide, we describe our analysis of geotagged content about *Indigenous communities* on Wikipedia. We chose Indigenous communities because the United States has a history of forced

resettlement of indigenous peoples to rural areas, with today 68% of Native Americans living on or near tribal lands, while 54% of Native Americans live in rural areas [4]. Further, Indigenous Americans are underrepresented in research on Internet usage [7, 8], and we wanted to increase awareness of representation of content created by and about these Americans.

2 METHODS

We constructed our datasets by crawling Wikipedia for geo-tagged content, generating different sets of pages by changing the seed page for the crawl. We leveraged Wikipedia lists that categorize pages based on criteria such as country, state, etc. For example, in order to find places with Native American majority populations, we started with https://en.wikipedia.org/wiki/List_of_U.S._communities_with_Native_American_majority_populations. We used different Wikipedia lists to create datasets consisting of pages from various states. We then retained only those pages which had geo-location and population information. In order to compare the content quality of articles about places with Native American majorities against rural and urban place based content, we then classified the pages into one of the four categories described below.

- Places with Indigenous majorities
- Rural places: Places with fewer than 2500 inhabitants
- Urban Clusters: Places with greater than 2500 but fewer than 50,000 inhabitants
- Urban Areas: Places with greater than 50,000 inhabitants

The bounds on population for each category are per the US Census Bureau's definitions of Urban, Urban Cluster and Rural places [3]. We restricted our dataset to states with significant Indigenous populations so as to provide a common ground for comparison. Table 1 summarizes the datasets.

Dataset	Number of Pages
Places with Native American Majority	401
Rural	4370
Urban Clusters	1498
Urban	97

Table 1: Dataset Summary

Once we decided the set of pages constituting each dataset, we then used the Wikimedia API [1] to download the article information. We downloaded page specific information, such as article length and title and also gathered the metadata associated with each revision of a page. We used the *user*

Permission to make digital or hard copies of part or all 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. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
COMPASS '20, June 15–17, 2020, , Ecuador
© 2020 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-7129-2/20/06.
<https://doi.org/10.1145/3378393.3402249>

Figure 1: Human and Bot Edit Distribution

Figure 2: Bot/Human Edit Distribution

property of the revision to determine if the creator of the edit was a human or a *bot*. A bot is an automated software agent which performs tasks such as citation cleanup and improving the formatting of pages. The names of such bot users names have the words “bot” or “awb” (AutoWikiBrowser) in their username.

3 RESULTS

We find significant differences in authorship across our datasets, in terms of the amount of bot generated content. Figure 1 separates bot edits from human edits for each of our datasets. While the contribution by bots is under 10% in Urban areas, it increases to over 30% in Indigenous communities. Prior work showed the significant increase in bot generated content from Urban to Rural and we find a similar contribution by bots to Indigenous geotagged pages [5].

Figure 2 shows that the Indigenous and Rural datasets have fewer pages with greater number of edits, whereas the opposite is observed for the Urban Cluster and Urban datasets.

Dataset	Average Article Length in Bytes
Places with Native American Majority	4508
Rural	5954
Urban Cluster	10306
Urban	34671

Table 2: Article Length Across Datasets

Both Figures 1 and 2 imply a low editor participation in the pages in Indigenous and Rural datasets.

We then looked at the length of the articles in each dataset. Table 2 shows that the average article length for the Urban dataset is several times that of the other datasets. This is further indication that the articles in the Indigenous and Rural datasets stand to improve in terms of amount of content and editor time devoted to them.

4 CONCLUSION

We find that articles about places with Native American majorities receive far less editor attention than pages about Urban Clusters or Urban areas on Wikipedia. They are maintained to a significant extent by bots. At the same time, there is growing awareness about the existence of a digital divide and its manifestations in Wikipedia. Over the years, there have been several initiatives, both within and outside of Wikipedia to contribute to pages concerning ethnic minorities. This, coupled with the fact that access to information has improved over time, leads us to believe that the amount of and quality of content about places with Indigenous majorities can improve over time.

ACKNOWLEDGMENTS

This work was supported by NSF Grant CNS-1563436

REFERENCES

- [1] Wikipedia API. 2019. API:Main Page. https://www.mediawiki.org/wiki/API:Main_page
- [2] Ofer Arazy and Oded Nov. 2010. Determinants of wikipedia quality: the roles of global and local contribution inequality. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*. ACM, 233–236.
- [3] US Census Bureau. 2019. <https://www2.census.gov/geo/pdfs/reference/GARM/Ch12GARM.pdf>
- [4] Sarah Dewees and Benjamin Marks. 2017. Twice invisible: Understanding rural Native America. *First Nations Development Institute, Research Note 2*, April (2017), 1–10.
- [5] Isaac L Johnson, Yilun Lin, Toby Jia-Jun Li, Andrew Hall, Aaron Halfaker, Johannes Schöning, and Brent Hecht. 2016. Not at home on the range: Peer production and the urban/rural divide. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 13–25.
- [6] Aniket Kittur and Robert E Kraut. 2008. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*. ACM, 37–46.
- [7] Christian Sandvig. 2012. Connection at Ewiiapaayp Mountain. In *Race After the Internet*, Lisa Nakamura and P Chow-White (Eds.). Routledge, New York, NY, Book section 8, 168–200.
- [8] Morgan Vigil, Matthew Rantanen, and Elizabeth Belding. 2015. A first look at tribal web traffic. In *Proceedings of the 24th International Conference on World Wide Web*. Association for Computing Machinery, New York, NY, 1155–1165.