Information flow on COVID-19 over Wikipedia: A case study of 11 languages

Changwook Jung
changwook.jung@kaist.ac.kr
KAIST
Institute for Basic Science
Daejeon, South Korea

Inho Hong
hong@mpib-berlin.mpg.de
Max Planck Institute for Human Development
Berlin, Germany

Diego Sáez-Trumper
dsaez-trumper@acm.org
Wikimedia Foundation
San Francisco, CA, USA

Damin Lee
damini@postech.ac.kr
POSTECH
Pohang, South Korea

Jaehyeon Myung
mjhbest@kaist.ac.kr
KAIST
Daejeon, South Korea

Danu Kim
danu@kaist.ac.kr
KAIST
Daejeon, South Korea

Jinhyuk Yun
jinhyuk.yun@ssu.ac.kr
Soongsil University
Seoul, South Korea

Woo-Sung Jung
wsjung@postech.ac.kr
POSTECH
Pohang, South Korea

Meeyoung Cha
mcha@ibs.re.kr
Institute for Basic Science
KAIST
Daejeon, South Korea

ABSTRACT

Wikipedia has been a critical information source during the COVID-19 pandemic. Analyzing how information is created, edited, and viewed on this platform can help gain new insights for risk communication strategies for the next pandemic. Here, we study the content editor and viewer patterns on the COVID-19 related documents on Wikipedia using a near-complete dataset gathered of 11 languages over 238 days in 2020. Based on the analysis of the daily access and edit logs on the identified Wikipedia pages, we discuss how the regional and cultural closeness factors affect information demand and supply.

ACM Reference Format:
Changwook Jung, Inho Hong, Diego Sáez-Trumper, Damin Lee, Jaehyeon Myung, Danu Kim, Jinhyuk Yun, Woo-Sung Jung, and Meeyoung Cha. 2021. Information flow on COVID-19 over Wikipedia: A case study of 11 languages. In *Companion Proceedings of the Web Conference 2021 (WWW '21 Companion), April 19–23, 2021, Ljubljana, Slovenia.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3442442.3452352

1 INTRODUCTION

With the coronavirus pandemic, we are facing social phenomena that we have never been experienced before. Among them, the spread of rumors and fake news related to coronavirus caused serious social problems. The propagation of misinformation, which spreads similarly to an epidemic, is called infodemic [4]. Infodemic adversely affects the spread of infectious diseases by focusing people's attention on misinformation, and it hurts not only public health but also society as a whole. To reduce the damage of infodemic, blocking misinformation delivery early by providing accurate information is necessary. Wikipedia is an online encyclopedia built with collective intelligence and has faithfully served as a reliable information provider during the COVID-19 pandemic [1]. However, documents related to coronavirus written in 317 language services show language-specific writing patterns and information consumption. To mitigate the infodemics, we selected 11 languages to study how the necessary public health-related information was generated and consumed.

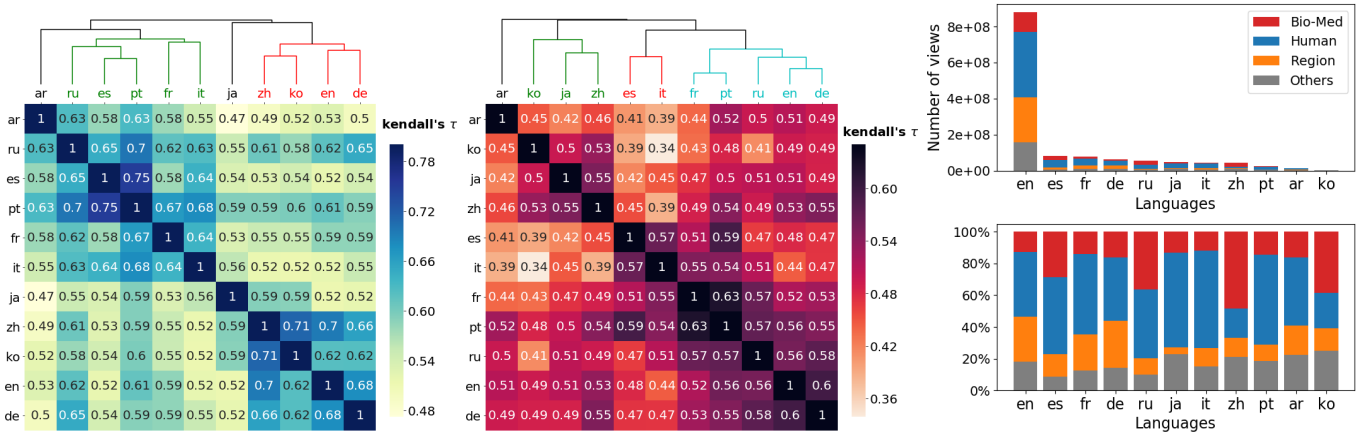
We use the near-complete data collected under a novel collection strategy in [2], an open dataset that addresses and resolve the problem of the keyword-based search. This new COVID-19 Wikipedia data is accompanied by the view count, edit count, network, and creation timestamp of documents between January 1 and November 20, 2020.

As a case study, we chose to examine patterns seen in 11 language projects out of 317 languages provided by Wikipedia: English(en), Spanish(es), Portuguese(pt), Italian(it), German(de), French(fr), Russian(ru), Arabic(ar), Chinese(zh), Korean(ko), and Japanese(ja). We selected English, Chinese, Korean, and Italian with their matching countries; the U.S., China, Korea, and Italy, which had shown early regional infection stages. The other chosen languages cover diverse populations, cultures, and continents.

Each Wikipedia page item in the used dataset had topical categorization of the following: Bio-Med, Region, People, and Others. Such categorization makes it easy to analyze how the interests of editors and viewers change over time. Among them, the Bio-Med topic contains direct information about the coronavirus (e.g., Wikipedia page titled 'COVID-19,' 'COVID-19 pandemic,' and 'SARS-CoV-2'). The Region topic includes documents containing local information, such as the 'COVID-19 pandemic in the United States.' The People topic included those infected with the virus as well as the relevant politicians and medical doctors. The rest of the documents were

WWW '21 Companion, April 19–23, 2021, Ljubljana, Slovenia

ACM ISBN 978-1-4503-8313-4/21/04.
https://doi.org/10.1145/3442442.3452352



(a) Rank correlation heat map by view count **(b) Rank correlation heat map by edit count** **(c) Total number of views for each category**

Figure 1: Information supply and demand on COVID-19 seen over 11 language projects. (a) and (b) display the hierarchically clustered rank correlation matrix based on view and edit count of all-language-items, where clusters indicate similarity in readers' interests (view) or authors' interests (edit) by language. Both heat maps indicate that the lighter the color, the lower the correlation, and the darker the color, the higher the correlation. (c) illustrate the total number of items and views for four main topical categories: bio-med, human, region, and others. The right figures show the category ratio of left figures.

classified as Others. For example, the '2020 stock market crash' item is classified in the Other category.

The subset of documents that appear in all 11 languages can be used to compare public interests. Based on the view and edit logs, we obtained a language-specific ranked list and measured their similarity by Kendall's rank correction coefficient. The result is then used as a metric to apply the UPGMA[3] algorithm, which clusters the most similar item pairs in an agglomerative way. Fig 1a is a heat map with a dendrogram that shows the lists' similarities and clusters based on information consumption. In contrast, Fig 1b is based on information edit frequency.

Rank correlation coefficients are between 0 and 1 and they varies between 0.47 and 0.75 in view count rank lists. The coefficients by edit count shows values between 0.34 and 0.63.

We find that the Spanish and Portuguese adjoined, and apart from German, most European language lists are in the same group, while east Asian languages, such as Korean, Chinese, and Japanese, are attached. Japanese has least associated with all the other languages. Japanese is out of the Korean and Chinese cluster in dendrogram based on view count, while more attached with Chinese and Korean in edit count based cluster. The all-language-item ranked lists appear to be related to the country's cultural, geographic distance associated with each language.

Finally, the number of coronavirus pandemic related documents and the amount of access were examined by category (Fig 1c). Although the number of items in the Bio-Med category is just three, the view count is relatively high, indicating their high demand.

These findings demonstrate value in looking into language-wise information supply and demand patterns on Wikipedia. While omitted here, we also observe cultural and geographical connections by grouping languages with similar interests in the information. By identifying imbalances in information delivery speed and information consumption, one may decide which information delivery strategy can help better information flow for future pandemics.

We will analyze the difference in information generation and consumption tendency by measuring the information generation speed and pattern in each document by Wikipedia language service more precisely through the information amount. Also by analyzing the hidden features and associations in the information delivery pattern for each language, our next study will aim to establish a basis of the information provide strategies for each language, culture, and country.

REFERENCES

[1] Giovanni Colavizza. 2020. COVID-19 research in Wikipedia. *Quantitative Science Studies* 1, 4 (2020), 1349–1380. https://doi.org/10.1162/qss_a_00080
[2] Changwook Jung, Diego Saez-Trumper, Inho Hong, and Meeyoung Cha. 2020. COVID-19 Wikipedia data. https://doi.org/10.6084/m9.figshare.13239515.v4
[3] Robert R Sokal. 1958. A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.* 38 (1958), 1409–1438.
[4] John Zarocostas. 2020. How to fight an infodemic. *The lancet* 395, 10225 (2020), 676.