Wiki Workshop (11th edition) – June 20, 2024

Estimating Gender Completeness in Wikipedia

Hrishikesh Patel, Tianwa Chen, Ivano Bongiovanni, Gianluca Demartini
The University of Queensland, Australia

Abstract

Gender imbalance in Wikipedia content is a known challenge which the editor community is actively addressing. The aim of our work is to provide the Wikipedia community with instruments to estimate the magnitude of the problem for different entity types (also known as classes) in Wikipedia. To this end, we apply class completeness estimation methods based on the gender attribute. Our results show not only which gender for different sub-classes of Person is more prevalent in Wikipedia, but also an idea of how complete the coverage is for difference genders and sub-classes of Person.

Keywords: gender; bias; wikipedia; statistical estimation; completeness;

1 Introduction

Wikipedia naturally grows and evolve over time. This happens while having human editors focussing on certain parts of the project instead of others. While the ability for editors to decide what to contribute comes with the advantage of flexibility, it may result in biased content where, for example, one gender is better represented than others. An example of this is the number of male astronauts as compared to the number of female astronauts (76 female out of 630 astronauts in Wikipedia, as of the submission of this paper).

The editor community is actively addressing this (Langrock and González-Bailón, 2022). Previous studies have shown how the editor population is also unbalanced from a gender point of view. (Antin et al., 2011) shows how the majority of editors, about 80%, are male. Another important aspect to understand gender representation is that of measuring how many persons of a certain gender are represented by an article in Wikipedia. To this end, we look at instances of sub-classes of the class Person (e.g., astronauts). Such instances are represented by a single Wikipedia article (e.g, `https://en.wikipedia.org/wiki/Samantha_Cristoforetti`). The first step is that of counting how many instances (i.e., person) of a certain gender there are in such a class (e.g., astronaut) and make observations (e.g., 554 male astronauts and 76 female astronauts).

The more interesting step after counting entities is that of understanding how well represented each of the genders are in Wikipedia. To do this we would need to measure how many male/female astronauts there *should be* in Wikipedia, assuming that the class is not yet completely represented in Wikipedia (e.g., because of a focus of the editor community on different parts of Wikipedia).

To close the gender gap in Wikimedia content it is first critical to be able to measure it. To this end, in this paper we estimate the cardinality of sub-classes of Person based on gender. Similar to our approach, previous work by (Luggen et al., 2019) has used statistical estimators for class cardinality using Wikidata edit history in a capture/recapture setup. In a similar fashion, in our work we apply such estimators but using the Wikipedia edit history and also taking a gender-based approach to it. This allows us to not only estimate the cardinality of a class (e.g., female astronauts), but also, by comparing the estimated size with the number of male/female instances currently present in Wikipedia, to measure the *completeness* of each class for different genders (e.g., male astronauts are 95% complete while female astronauts are 94% complete), which is our research question.

2 Related Work

Previous research (Luggen et al., 2019) has looked at how to use statistical estimators to estimate class cardinality in Wikidata. They used the knowledge graph edit history as evidence for the estimators. In this paper we extend this approach by looking at attribute-specific cardinality estimations (e.g., How many female astronauts should be there? Do we have them all?) and beyond the Wikidata project.

In the area of crowdsourced databases, the problem of answering queries under the open world assumption has been studied in the past. Researchers have encountered the problem that popular entities are reported by crowd members more frequently than "tail" (i.e., unpopular) entities, thus making it difficult to complete the answer set (and, in our case, to estimate the class cardinality). The approach followed by (Trushkowsky et al., 2014) has looked at using statistical estimators to understand how

far from the complete set the incrementally constructed query answer set currently is. We apply these methods to understand how complete Wikipedia is.

3 Data Collection

Our aim is to generate a dataset that enables us to estimate gender-based class cardinality first and completeness levels next. To this end, we first collected the list of sub-classes (e.g., Artist, Astronaut, Monarch, etc.) of the class 'Person' from DBpedia[1] as well as the list of all entities assigned to each of these sub-classes. For each of the collected entities, we then retrieved its Wikipedia article edit history from the MediaWiki API for the period from Jan 2019 to Dec 2023 to be used as "capturing" events for the statistical estimators. In total, our dataset contains 121,535 entities over 34 classes and a total of 4,896,299 edits. The size of the dataset is 1.2GB and was processed on a 32GB RAM server.

4 Name-based Gender Estimation

Based on previous work by (Van Buskirk et al., 2023), we first estimate the gender of persons described by the Wikipedia articles. Using the dataset described above, we perform name-based gender classification for each of the collected entities of type 'Person'. This enables us to make initial observations about the gender distribution of persons having an article in Wikipedia for different sub-classes of Person.

Figure 1 shows the results of this classification task[2]. We can observe how, unsurprisingly, certain sub-classes of Person are less gender balanced than others (e.g., 'Economist' being mostly male and 'Model' being mostly female).

5 Gender-based Class Cardinality Estimation

Once we have classified the gender of the people with a Wikipedia article, we can now proceed with the estimation of the cardinality of each sub-class of Person (i.e., how many Astronauts should we have in Wikipedia) using the edit history of their articles. In short, these methods make use of samples of entities from a population (e.g., Astronauts) to estimate the size of the population. The capture/recapture methods we use, originally designed in computational ecology (e.g., capturing and tagging lions in the Savanna to estimate the size of the entire population of lions), require a concept of 'sampling' (i.e., capturing samples of the population over time). In our setting, we make use of the edit of a Wikipedia article as the sampling event. More popular entities will receive more edits (and thus will be sampled more often) than less popular ones. With this data, the more samples we have the more accurate the estimators will be, eventually converging to the true value of the population size. The other good aspect of these estimators is that they provide a confidence score that tell us how accurate the estimation is. More details on the statistical estimation methods we use can be found in (Luggen et al., 2019). Some example population size estimation results are presented in Figure 2.

These methods allow us to compute how many entities of a certain type there *should* be in Wikipedia and, with that, compute an estimated completeness level by comparing the estimated cardinality with the current number of entities of that class in Wikipedia.

6 Results and Conclusions

Overall, we observed high estimated completeness levels. This is consistent with the generally perceived high quality of Wikipedia content. Our estimates show that 16 sub-classes have higher completion rate for male and 13 have a higher completion rate for female. Possible limitations include the use of DBpedia sub-classes of Person which may be imperfect, as well as the impossibility to assess estimation accuracy due to the lack of a ground truth.

[1] https://www.dbpedia.org/about/

[2] The approach would not classify a name if its confidence is low and thus we keep a column for undefined gender in Fig. 1

References

[Antin et al.2011] Judd Antin, Raymond Yee, Coye Cheshire, and Oded Nov. 2011. Gender differences in wikipedia editing. In *Proceedings of the 7th international symposium on wikis and open collaboration*, pages 11–14.

[Langrock and González-Bailón2022] Isabelle Langrock and Sandra González-Bailón. 2022. The gender divide in wikipedia: Quantifying and assessing the impact of two feminist interventions. *Journal of Communication*, 72(3):297–321.

[Luggen et al.2019] Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. 2019. Non-parametric class completeness estimators for collaborative knowledge graphs—the case of wikidata. In *The Semantic Web–ISWC 2019: 18th International Semantic Web Conference*, pages 453–469. Springer.

[Trushkowsky et al.2014] Beth Trushkowsky, Tim Kraska, Michael J Franklin, Purnamrita Sarkar, and Venketaram Ramachandran. 2014. Crowdsourcing enumeration queries: Estimators and interfaces. *IEEE TKDE*, 27(7):1796–1809.

[Van Buskirk et al.2023] Ian Van Buskirk, Aaron Clauset, and Daniel B Larremore. 2023. An open-source cultural consensus approach to name-based gender classification. In *AAAI ICWSM*, volume 17, pages 866–877.