Measuring the Gender Gap: Attribute-based Class Completeness Estimation

Gianluca Demartini
The University of Queensland, Australia

Abstract

"If you can not measure it, you can not improve it.", Lord Kelvin.

Understanding the magnitude of a problem is the first step towards solving it. Gender imbalance in Wikipedia is a known problem which the editor community is actively addressing [LG 2022]. The aim of this project 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. In this research project, we propose to design and evaluate novel class completeness estimation methods based on entity attributes such as gender. This would allow us to provide a way to measure the gender gap in collections evolving over time, such as, for example, Wikipedia or Wikidata. By focusing on Wikipedia, we will address open research questions such as i) how to incorporate changes in gender balance over time into the statistical estimators currently available from the literature and ii) how to incorporate the notability criteria when providing information to the community of editors about how many entities of a certain gender are currently missing.

Introduction

Successful crowdsourcing projects like Wikipedia and Wikidata naturally grow 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 to 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 (73 out of 574, https://en.wikipedia.org/wiki/List_of_female_astronauts https://en.wikipedia.org/wiki/List_of_space_travelers_by_name) in Wikipedia.

There are possible viable approaches to address this issue. For example, the editor community may decide to stop adding new male astronauts to the project to allow for content about female astronauts to catch up. Alternatively, the community may decide to target the actual gender distribution in the profession. In any case, the choice of how to address the gender gap remains an editor community decision.

To close the gap (the gender gap in Wikimedia project content, in our case) it is first critical to be able to *measure* it. While it is up to the editor community to make decisions on how to best balance content by prioritizing editorial

focus, in this research project we instead aim at empowering editors in their decisions by providing them with relevant information about the current gender balance across categories in Wikipedia. This aligns with the 2030 Wikimedia Strategic Direction as our contribution enables the platform and the community to collect “knowledge that fully represents human diversity”.

Research Contribution. Rather than deciding how to deal with gender unbalanced content, the aim of this research is to automatically identify underrepresented classes by quantifying and measuring the expected size of a class in order to empower the community in taking decisions and setting editorial priorities. This is possible by making use of the edit history for Wikipedia articles. While this can be done by adapting previous research, there are a number of open research questions (RQs) that need to be addressed before these methods can be applied effectively to Wikipedia. These RQs include the following:

- When we consider open classes that grow over time and gender balance ratios that change over time, how can we provide accurate completeness estimations?
- When not all entities are to be included (e.g., because of the notability criteria), how can we decide on the correct number of entities that should be included?

Related work

Previous research has already looked at the gender gap [FSS 2016], which is a particularly important issue across Wikimedia projects [RGJ+ 2020]. For example, authors of [AMS 2022] proposed methods to identify content gaps in crowdsourced knowledge graphs and found that Wikidata editors usually tend not to work on under-represented entities. While many attempts to address this issue exist, the key difference in the approach we propose to take is that we do not attempt to address the issue, but rather, to design effective methods and develop accurate instruments for the different Wikipedia user groups to be empowered and able to address the issue by themselves.

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 [TKF+ 2014] has looked at using statistical estimators to understand how far from the complete set the incrementally constructed query answer set currently is. Inspired from this approach, we recently proposed methods to estimate the class cardinality in crowdsourced knowledge graphs [LDC+ 2019].

Our previous research [LDC+ 2019] has looked at how to use statistical estimators to estimate class cardinality in Wikidata. We used the knowledge graph edit history as evidence for the estimators and to measure class completeness. In this project we plan to extend our approach by looking at attribute-specific cardinality estimations (e.g., How many women astronauts should be there? Do we have them all?) and beyond the Wikidata project. Specifically, we will focus on Wikipedia and will expand state-of-the-art methods to include aspects like open classes and changes in population size over time, as well as the notability criteria used by Wikipedia (see Methods section). To this end, we will develop novel class cardinality estimators for Wikidata categories based on articles’ edit history.

Methods

Our previously proposed method [LDC+ 2019] can estimate the cardinality of a class of entities based on repeated observation of instances of that class (e.g., from the edit history). Hence, this method can be used to answer questions such as “Does the knowledge base have a complete list of all female astronauts?”. Our techniques are derived from species estimation and data management and are applied to the scenario of collaborative editing, like in Wikimedia projects. We make use of entities observed in a project’s edit history as a proxy for observations in a capture/recapture study setup. This allows us to use estimators for species population (e.g., Jackknife Estimators [HF 1983]) to predict class cardinality [LKP+ 2005].

This approach can be generalized easily and can be applied to non-binary value attributes (e.g., non-binary genders, or age groups like, for example, counting how many astronauts in the age ranges 20-30, 30-40, and 40-50 there should be in Wikipedia) to estimate attributed-based class cardinality. In the following we present the dataset we plan to use in this project, and the structure of the proposed project activities aimed to develop accurate estimators as well as to address the open research question presented above.

Dataset.
We plan to deploy and evaluate the proposed methods on the English Wikipedia article edit history using Wikipedia categories as an initial proxy for entity classes (similarly to the approach followed to construct YAGO [SKW 2008]). This is different from [LDC+ 2019] where we used Wikidata classes and Wikidata entity edit history instead.

Task 1. Model of population growth over time.

To deal with open classes that keep growing over time and that are potentially unbounded, we plan to extend existing population estimators as used in [LDC+ 2019] by incorporating the notion of a population’s growth ratio. For example, in [DNR+ 2002] authors adapted population estimators by including the expected exponential growth of human cells affected by a virus. In this task, we plan to follow a similar approach to deal with open classes in Wikimedia projects. This would allow us to develop new class cardinality estimation methods that take into account the expected population growth which can be tracked over time in the dataset. Note, however, that the population growth ratio may not be stable over time and could be affected both by sporadic events as well as by long-term trends. This additional challenge will be addressed in the next task by running the growth ratio prediction models regularly over time.

Task 2. Concept Drift Management.
A problem related to open classes which grow over time at an indefinite rate is that of *concept drift* [GZB+ 2014]. While this is typically defined for supervised classification problems where training evidence for the target class changes over time (e.g., predicting consumer behavior during a pandemic), it can be mapped to our problem of estimating a class cardinality for class attributes that change over time.

Changes in data over time may happen at different rates: abruptly or incrementally. It is then critical to detect these changes and apply the relevant adjustments to the estimates made by our method. Most of the approaches designed to deal with concept drift in supervised machine learning model make use of “forgetting” mechanisms that can be translated into our setting as using a time window (e.g., a temporal interval) over the edit history data used to make population size

estimations or by means of a “fading factor” that gives more weight to recent observations. The faster the change in the population size, the shorter the time interval we want to consider data from. This allows us to focus on recent evidence to make population size estimates more accurate.

In machine learning it is often challenging (because of the computational cost) to re-train a supervised model when the underlying data changes. This has led to having a preference for incremental models that can adapt to changing data environments. In our context, however, it is computationally more efficient to re-train class size estimators in order to deal with changes in the data. In addition, because of the lower computational complexity, it is possible to consider ensembles of estimation models (e.g., combining those used in [LDC+ 2019]) that, aggregated together, could better deal with drifts in population trends over time.

In this project, the idea of concept drift can be mapped to the situation in which gender balance for a class changes over time. For example, the population of mathematicians has had different gender balances over history. If Wikipedia is to keep records of all mathematicians over a long period of time, then it is important to monitor changes in the population gender distribution to be aware if the currently represented entities reflect the actual balance in the population or not. By using methods from concept drift management, we expect to be able to develop novel statistical estimators that can, over time, adjust to changes in the population gender distribution.

Task 3. Estimator adjustments and evaluation.
Based on the results of the research conducted in Task 1 and 2, in this task we will perform an experimental evaluation of the extended class completeness estimation methods using the proposed dataset to validate their robustness when incorporating population growth models and concept drift approaches. We will use the models from [LDC+ 2019] as a baseline for comparison and measure error rates in the produced estimates over time in order to reach a conclusion on the added benefit of the extended approach. Beyond insights on the effectiveness of the methods, this will generate additional data about the gender gap in Wikipedia that the community can take on board in their decision making processes.

Task 4. Notability Ranking.
Not all entities of a certain class appear in Wikipedia due to the notability criteria (https://en.wikipedia.org/wiki/Wikipedia:Notability). When looking at the gender gap in Wikipedia, it is important to understand if the gender distribution in a class is also reflected in the notable entity distribution for that class. This is a complementary effort to that of attempting to measure gender-based class completeness. For example, while many cultures like in Russia see many female students choosing STEM disciplines thus resulting in a well gender-balanced population, less women end up becoming famous in their discipline (again, for example, due to a culture promoting male leaders) and showcasing achievements, receive awards, and “significant coverage” that would make them “notable” and thus more likely being the subject of a Wikipedia article.

Given the chosen dataset and approach, in this project we thus focus on gender-based class completeness estimation for notable entities as the method is based on the Wikipedia article edit history and thus only observes those entities. To compare the estimated gender balance among notable entities with the gender distribution of the general class population (i.e., including non-notable entities) we plan to use external data sources and make this information

available to the editor community together with the notable entities gender completeness estimates. This would allow a more informed community-driven decision-making process as they could decide whether to prioritize classes where there is a gender representation issue in notable entities but not in the general class population, thus potentially leading to Wikipedia as an instrument for fostering positive culture change.

In addition, in this task we aim to develop notability-based entity ranking methods (see, for example, [KSD+ 2010] on how to use Wikipedia and hyperlinks pointing to Wikipedia to do such ranking) that can sort entities based on an algorithmically computed notion of notability.

Rather than having an automated method making the subjective decision on whether an entity is deemed to be notable or not, we rather envision a human-in-the-loop approach where the role of the algorithm is to generate a notability-based ranking and then leave to the human editors the decision on whether to include or not certain entities in Wikipedia.

Task 5. Community Engagement.
An important non-technical task in this project is to obtain feedback from Wikipedia user groups. In this task we will engage with the community of Wikipedia editors with a specific focus on user groups currently looking at the gender content gap. The aim is to identify collaborators who can, on the one hand, provide early feedback to guide our research and, on the other hand, help us develop instruments which the community might be willing to adopt so that they could benefit from the estimates generated by our methods.

Task 6. Write up and Publication.
In this task the principal investigator will support the RA in writing up the methods and the experimental results into a submission for an academic conference publication. In case of acceptance, the work will be then presented at an academic research conference.

Expected output

We envision the following outputs:

- One scientific publication (e.g., at the International AAAI Conference On Web And Social Media, ICWSM). discussing the novel methods and an experimental evaluation of their effectiveness co-authored by the principal investigator and the research assistant. The primary intended audience for this output is the academic research community.
- A software instrument (e.g., a dashboard or a script depending on resource availability) that can be used to estimate gender-based class completeness over time in Wikipedia. The primary intended audience for this output is Wikimedia Foundation or Wikipedia users with technical skills who can run the instrument to generate insights for the community.
- Insights to inform decision making by the Wikipedia user community on which parts of the project to focus on. The primary intended audience for this output is the community of editors who would be empowered in taking data-driven editorial decisions.

Risks

Possible project risks include skilled staff availability, and low community engagement.

Staff: We plan to recruit project staff (i.e, the Research Assistant) among graduating students in the Master of Data Science degree at the University of Queensland (UQ). This will allow us to tap into a large pool of candidates (aprox. 100 graduates/year) with the right skills for the project. For the community

engagement role we plan to identify a relevant person from students in other UQ departments (i.e., beyond Computer Science).

Community Engagement: it is hard to predict how the editor community will perceive the proposed methods. We budgeted for a community engagement role to dedicate the necessary time to seek and listen to any community feedback and to incorporate it on board of our solution design to increase the likelihood of future acceptability and adoption.

Date: This project will start on July 1, 2023 and will conclude on June 30, 2024. Here is a time plan for the tasks in the project:

	M1-2	M3-4	M5-6	M7-8	M9-10	M11-12
T1						
T2						
T3						
T4						
T5						
T6						

Community impact plan

We will seek to impact audiences beyond researchers and academics by means of Task 5 and the community engagement role envisioned in the project budget.
We plan to work with different Wikipedia user groups (e.g., https://meta.wikimedia.org/wiki/WikiWomen%27s_User_Group) which focus on the gender gap issue to seek their feedback on our proposed instrument design and make the expected output more adoptable and impactful.

To this end, the community engagement staff member will start at the beginning of the project (i.e., month 1-2, see also Gantt chart above) to seek early feedback on the overall project idea and solution design. Then, once the solution development has reached a good maturity level, this staff member will engage again with the user groups to understand the potential for adoption and seek to collect any further additional requirement that could increase it.

Evaluation

We will consider the following research success factors and metrics:

- Effectiveness gains over baseline methods. This will be the result of the experimental evaluation planned in Task 3 and will be measured following a similar experimental setup as in [LDC+ 2019] and using methods in that work as a baseline for comparison.
- Community adoption. We aim for the developed instruments to be used by the Wikipedia editor community to inform their editorial decision making processes. We will track usage of the available instruments (see Output section) and measure it over time. We would rather consider success a prolonged use by few users, rather than a high absolute number of distinct users over a short period of time.

Budget

We budgeted for a total cost of approximately $20,000 USD. This is mostly to cover a casual research assistant (RA) for 270 hours to work on the technical project tasks. We will seek to fill the RA role from recent UQ graduates who have skills in data science and software development.

In addition, we budgeted for 50 hours of a community relationship manager who will be responsible for the community engagement aspects of the project (Task 5) and seek community feedback to make sure the project outcome can reach the desired impact.

Finally, we budgeted for the costs to publish and attend a scientific conference to present the research conducted in the project (e.g., by means of an accepted paper presentation). Please, refer to the linked spreadsheet for a full cost breakdown.

Response to reviewers and meta-reviewers

- [Redacted]

References

[AMS 2022] Abián, D., Meroño-Peñuela, A., & Simperl, E. (2022, October). An Analysis of Content Gaps Versus User Needs in the Wikidata Knowledge Graph. In The Semantic Web–ISWC 2022: 21st International Semantic Web Conference, Virtual Event, October 23–27, 2022, Proceedings (pp. 354-374). Cham: Springer International Publishing.

[DNR+ 2002] Drummond, A. J., Nicholls, G. K., Rodrigo, A. G., & Solomon, W. (2002). Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics, 161(3), 1307-1320.

[FSS 2016] Farzan, R., Savage, S., & Saviaga, C. F. (2016). Bring on board new enthusiasts! A case study of impact of Wikipedia art+ feminism edit-a-thon events on newcomers. In Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I 8 (pp. 24-40). Springer International Publishing.

[GZB+ 2014] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM computing surveys (CSUR), 46(4), 1-37.

[HF 1983] Heltshe, J.F., Forrester, N.E.: Estimating species richness using the jackknife procedure. Biometrics pp. 1–11 (1983)

[KSD+ 2010] Kaptein, R., Serdyukov, P., De Vries, A., & Kamps, J. (2010, October). Entity ranking using Wikipedia as a pivot. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 69-78).

[LKP+ 2005] Lancia, R. A., Kendall, W. L., Pollock, K. H., & Nichols, J. D. (2005). Estimating the number of animals in wildlife populations.

[LG 2022] Langrock, I., & González-Bailón, S. (2022). The Gender Divide in Wikipedia: Quantifying and Assessing the Impact of Two Feminist Interventions. Journal of Communication, 72(3), 297-321.

[LDC+ 2019] Michael Luggen, Djellel Difallah, Cristina Sarasua, Gianluca Demartini, and Philippe Cudré-Mauroux. Non-Parametric Class Completeness Estimators for Collaborative Knowledge Graphs. In: The International Semantic Web Conference (ISWC 2019 - Research Track). Auckland, New Zealand, October 2019.

[RGJ+ 2020] Redi, M., Gerlach, M., Johnson, I., Morgan, J., & Zia, L. (2020). A taxonomy of knowledge gaps for wikimedia projects (second draft). arXiv preprint arXiv:2008.12314.

[SKW 2008] Suchanek, F. M., Kasneci, G., & Weikum, G. (2008). Yago: A large ontology from wikipedia and wordnet. Journal of Web Semantics, 6(3), 203-217.

[TKF+ 2014] Trushkowsky, B., Kraska, T., Franklin, M. J., Sarkar, P., & Ramachandran, V. (2014). Crowdsourcing enumeration queries: Estimators and interfaces. IEEE Transactions on Knowledge and Data Engineering, 27(7), 1796-1809.