Generative Hypergraph Models for Understanding Wikipedia Users

Priyanka Sinha
Docyt, India

Dilys Thomas
Tata Consultancy Services Limited, India

Pabitra Mitra
Indian Institute of Technology Kharagpur
India

1 ABSTRACT

We propose to develop generative hypergraph models to understand Wikipedia users' behavior patterns. These models can estimate growth across various Wikimedia sections and inform strategies for improving user engagement globally.

2 INTRODUCTION

This research addresses key challenges in computational organizational behavior through the lens of Wikimedia contributors.

2.1 Universal Behavior Modeling

A significant challenge in computational organizational science is obtaining sufficient enterprise data to build machine learning models of behavior, personality, and group dynamics. We aim to build a universal behavior model based on the observation that organizational norms can be derived from dialogue traces. With the growing availability of dialogue data in enterprises, we can develop more robust behavioral models.

2.2 Significance for Wikimedia

This research is valuable for Wikimedia because:

- Wikimedia contributors form a global, internet-scale remote social enterprise
- Understanding contributor characteristics can inform design decisions for better engagement
- While English-language Wikimedia has high engagement from developed nations, participation remains lower in other languages and cultures
- Low contributor engagement in certain cultures, geographies, and languages falls below current internet access coverage, indicating engagement challenges

Wikimedia provides an invaluable source for understanding contributor personality and group behavior at scale. In large organizations, regular interaction between employees and psychologically trained personnel is infeasible due to scale and cultural diversity, making computational techniques more practical. This research is essential for making Wikimedia more inclusive globally, particularly important as Wikimedia content informs AI training, where lack of inclusion can lead to biased and irresponsible decision-making systems [9].

3 RELATED WORK

As authors, we have conducted prior work on Wikipedia data using DBpedia Spotlight [8, 11]. We have also published several related works in understanding behavior [10, 12–15].

4 METHODS

4.1 Data Collection

We will collect temporal user activity traces from Wikipedia Data Dumps [1], specifically:

- Logging database XML files (blocks, protection, deletion, uploads)
- File format: wikiname-YYYYMMDD-pages-logging.xml.gz

These will be converted into MySQL using existing Wikipedia tools. Our analysis will cover users across multiple languages beyond English. Data will be obtained, stored, and processed on a cloud server. No surveys will be conducted.

4.2 Analysis Approach

From the logs, we will:

- Represent user/IP address activity as a hypergraph where weighted edges represent the number of edited pages, and nodes represent users/IP addresses
- Develop temporal hypergraphs incorporating timestamps
- Extend the work in [4] to create generative models for weighted multi-edges and temporal hypergraphs
- Model growth patterns across different Wikipedia sections or language versions

Additionally, we will download talk page history dumps (wikiname-YYYYMMDD-stub-meta-history.xml.gz) to analyze psycholinguistic patterns in user discussions. These will also be represented as weighted multi-edged hypergraphs. By leveraging existing resources like lists of Wikipedians by edit count [2, 3], our generative models can provide insights on future growth patterns across various language communities.

5 EXPECTED OUTPUT

Our deliverables will include:

- Scientific publications targeted at venues such as ACM WebConf, ACM TKDD and Wiki Workshop
- Open-source code repository on GitHub

The primary audience will be web researchers and those engaged in web and user mining.

6 RISKS

The primary risk is computational resource limitations and intractability of hypergraphs [5, 6]. To mitigate this, we will:

- Optimize algorithmic efficiency
- Adjust cloud configuration as needed

7 EVALUATION

Success will be measured by:

- Peer-reviewed publication in conferences or journals with archival proceedings
- Public release of source code

8 BUDGET

Total requested budget: $35,284.32 USD [7]

REFERENCES

[1] [n. d.]. Wikimedia Downloads. https://dumps.wikimedia.org/backup-index.html

[2] [n. d.]. Wikipedia:List of Wikipedians by number of edits. https://en.wikipedia.org/wiki/Wikipedia:List_of_Wikipedians_by_number_of_edits

[3] [n. d.]. Wikipedia:WikiProject India/List of Indian Wikipedians by number of edits. https://en.wikipedia.org/wiki/Wikipedia:WikiProject_India/List_of_Indian_Wikipedians_by_number_of_edits

[4] Minyoung Choe, Jihoon Ko, Taehyung Kwon, Kijung Shin, and Christos Faloutsos. 2025. Kronecker Generative Models for Power Law Patterns in Real-World Hypergraphs. In *THE WEB CONFERENCE 2025*. https://openreview.net/forum?id=N73Yz5SQK9

[5] Vignesh Ganapathy, Dilys Thomas, Tomas Feder, Hector Garcia-Molina, and Rajeev Motwani. 2011. Distributing data for secure database services. In *Proceedings of the 4th International Workshop on Privacy and Anonymity in the Information Society* (Uppsala, Sweden) *(PAIS '11)*. Association for Computing Machinery, New York, NY, USA, Article 8, 10 pages. https://doi.org/10.1145/1971690.1971698

[6] George Karypis and Vipin Kumar. 1999. Multilevel k-way hypergraph partitioning. In *Proceedings of the 36th*

Annual ACM/IEEE Design Automation Conference (New Orleans, Louisiana, USA) *(DAC '99)*. Association for Computing Machinery, New York, NY, USA, 343–348. https://doi.org/10.1145/309847.309954

[7] Pabitra Mitra Priyanka Sinha, Dilys Thomas. 2025. Research Fund Budget. https://docs.google.com/spreadsheets/d/1JPiRcI5YG6wRmtNPD1jeRN6OqHn86YCPBYX4WYhIZgc/edit?gid=0#gid=0

[8] Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie (Eds.). 2015. *Proceedings, 5th Workshop on Making Sense of Microposts (#Microposts2015): Big things come in small packages, Florence, Italy, 18th of May 2015*.

[9] Priyanka Sinha. 2024. Researcher perspective of OPT-OUT. In *Internet Architecture Board (IAB): Workshop on AI-Control*. https://datatracker.ietf.org/doc/slides-aicontrolws-researcher-perspective-of-opt-out/

[10] Priyanka Sinha, Michael Ackermann, Pabitra Mitra, Arvind Singh, and Amit Kumar Agrawal. 2021. Characterizing the IETF through its consensus mechanisms. In *Internet Architecture Board (IAB): Show me the numbers: Workshop on Analyzing IETF Data (AID)*. https://www.iab.org/wp-content/IAB-uploads/2021/11/Sinha.pdf

[11] Priyanka Sinha and Biswanath Barik. 2015. Named Entity Extraction and Linking in #Microposts. In *5th Workshop on Making Sense of Microposts (#Microposts2015) collocated with WWW 2015*. Florence, Italy.

[12] Priyanka Sinha, Anirban Dutta Choudhury, and Amit Kumar Agrawal. 2014. Sentiment Analysis of Wimbledon Tweets. In *4th Workshop on Making Sense of Microposts (#Microposts2014) collocated with WWW 2014*. 51–52. http://ceur-ws.org/Vol-1141/paper_10.pdf

[13] Priyanka Sinha, Lipika Dey, Pabitra Mitra, and Dilys Thomas. 2020. A Hierarchical Clustering Algorithm for Characterizing Social Media Users. In *Companion Proceedings of the Web Conference 2020* (Taipei, Taiwan) *(WWW '20)*. Association for Computing Machinery, New York, NY, USA, 353–362. https://doi.org/10.1145/3366424.3383296

[14] Priyanka Sinha, Pabitra Mitra, Antonio Anastasio Bruto da Costa, and Nikolaos Kekatos. 2021. Explaining Outcomes of Multi-Party Dialogues using Causal Learning. arXiv:2105.00944 [cs.AI] https://arxiv.org/abs/2105.00944

[15] Priyanka Sinha, Ritu Patel, Pabitra Mitra, Dilys Thomas, and Lipika Dey. 2022. Mining Homophilic Groups of Users using Edge Attributed Node Embedding from Enterprise Social Networks. In *Companion Proceedings of the Web Conference 2022* (Virtual Event, Lyon, France) *(WWW '22)*. Association for Computing Machinery, New York, NY, USA, 1139–1147. https://doi.org/10.1145/3487553.3524726