Edit History Growth Rate Analysis

Leonhard Horstmeyer
Lakat♠, CSH⋆

Jonathan Kung
DFG ♦

Craig Calcaterra
DFG ♦, MSU♣

Benjamin Bose
Uni of Edinburgh

Kuzma Kudim
PHP dev

♠ Lakat - Permissionless ScientificPublishing, ♦ DAO Governance Framework, ⋆ Complexity Science Hub
♣ Metro State University

Abstract

We propose to investigate the DAG (Directed Acyclic Graph) growth rate of the edit history given the wikipedia data model and contrast this against more fine-grained version control models. We also investigate the data composability and reusability that is accompanied by a finer content resolution. Our methodology includes analyzing article histories using API endpoints and employing similarity metrics like Jaccard similarity and Kullback-Leibler divergence. The research will contribute insights into Wikipedia's data sustainability and content reusability, targeting the Wikimedia community and data science researchers.

Introduction

We would like to conduct research into the composability and modularity of the wiki data model. The mediawiki php-scripts and the layout of the wiki Entity-Relationship-Diagram reveal a versatile and performant software that at its core functions as a version control system with snapshots of each edit. The result is a database that grows linearly with the amount of edits per article. We would like to study this growth rate in depth and contrast that to alternative more granular data models with smaller content units such as sections. The higher granularity allows for more composability and reusability. A section for instance can be reused for the next edit or even for another article. To that end we propose to also study the extent to which wikipedia edits are reusable by looking at the content overlaps between edits or across different articles.

We believe that the wikimedia community will profit from this study in several ways: Data is developing as one of the main resources of this century. Investigating its growth rate and contrasting it with more granular options provides insights and opens up the possibility for a more data-sustainable future. With the rise of AI and bot-edits we expect that there will be a rise in data volume. Furthermore, granularity may also affect the process of merging pages, the identification of conflictual information and the ability to store content in a decentralized manner.

Date: We propose to start on June 1, 2024 and conclude by November 30, 2024.

Related work

A valuable survey of finding duplicate and contradictory information on Wikipedia has been conducted by Arazy et. al [1]. Moreover Urdaneta et. al. [2] have investigated the workload associated with storing wikipedia on a peer-to-peer network. In [3] a composable, mergeable and decentralized data model for wiki-like systems is discussed. Furthermore, some ideas about fine-grained version control systems are discussed in [4].

Methods

We are planning to use the available api endpoints to download a large sample of article histories and investigate which influence the modularization would have on the storage size. For instance, we would take a section as the smallest content unit and investigate the average space saved by updating just the section rather than snapshotting the entire article for each edit. We use python and php. For studying the reusability we will look at various similarity metrics, most prominently the Jaccard similarity, but also the Kullback-Leibler divergence.

Expected output

We expect four key outputs:

- Insights into the growth rate of the edit history of wikis.
- Comparison to various more fine-grained approaches to version control.
- Statistical analysis of the findings and dependence on article size, language and topic field
- Quantitative analysis of the reusability of a more granular data model. The output will look at the similarity between edits and across articles.

The primary audience for the analysis is the wiki community itself and researchers who are interested in quantitative studies of composability and reusability of the largest open knowledge database in human history.

Risks

As with any kind of quantitative research there is a possibility that the data is not as conclusive as was anticipated. We are willing to take this into account. We do not see any infringement of people's rights.

Community impact plan

This project would greatly profit from contributors with insights into data science and an understanding of the wikimedia data model. We would like to reach out to the Wikimedia volunteer developers. Some of us are members in the Wikimedia discord server and the Wikimedia Hackathon Telegram group. We would like to present our project there as well and furthermore would like to involve the community in both the process of research as well as the dissemination of findings.

Evaluation

Our research can be evaluated on two grounds. First on a quantitative basis, regarding the statistically significant results, the publication thereof and the soundness of the methods deployed therein. Secondly the project may be evaluated by the engagement of the Wikimedia community and the potential ramifications for projects emerging out of the research, such as a hackathon, a discussion about testing implementations of fine-grained version control systems for wikis.

Budget

We envisage that the salary is split between compensation for long hours of coding, data analysis and data cleaning, i.e. salary, and possibly open access publishing costs. We are a team of 5 working for 6 months, so we suggest $15000 for salary and $2000 for open access publishing fees. We would prefer however to publish without predatory fees.

Prior contributions

We are all quite passionate about data, protocols, quantitative and network analysis. Horstmeyer has conducted research in the field of network systems, analyzing metrics for the knowledge graphs of arxiv.org [5] and network

dynamics [6] [7] and higher-order connectivities [8]. Calcaterra has been investigating Organizational and Societal Structures [9] and Network-based metrics for reputation systems [10][11]. Kung and Calcaterra are both working on protocol research and DAOs, investigating metrics for reputation systems on DAGs such as git or blockchains [12]. That research is coordinated through a mediawiki [13]. Horstmeyer has also worked on a decentralized version-control protocol[14]. Bose is a researcher in cosmology and Kuzam a php and mediawiki-extension developer.

References

[1] Weissman, Sarah, Samet Ayhan, Joshua Bradley, and Jimmy Lin. "Identifying duplicate and contradictory information in wikipedia." In *Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries*, pp. 57-60. 2015.

[2] Urdaneta, Guido, Guillaume Pierre, and Maarten Van Steen. "Wikipedia workload analysis for decentralized hosting." *Computer Networks* 53, no. 11 (2009): 1830-1845.

[3] Horstmeyer, Leonhard. "Lakat: An open and permissionless architecture for continuous integration academic publishing." *arXiv preprint arXiv:2306.09298* (2023).

[4] Zolkifli, Nazatul Nurlisa, Amir Ngah, and Aziz Deraman. "Version control system: A review." *Procedia Computer Science* 135 (2018): 408-415.

[5] Hellmann, Jennifer, Leonhard Horstmeyer, Lin Li, Anna Olson, and Stefan Pfenninger. "Network Analysis of Interdisciplinary Research in the Physical Sciences." (2014).

[6] Horstmeyer, Leonhard, Tuan Minh Pham, Jan Korbel, and Stefan Thurner. "Predicting collapse of adaptive networked systems without knowing the network." *Scientific Reports* 10, no. 1 (2020): 1223.

[7] Horstmeyer, Leonhard, Christian Kuehn, and Stefan Thurner. "Network topology near criticality in adaptive epidemics." *Physical Review E* 98, no. 4 (2018): 042313.

[8] Horstmeyer, Leonhard, and Christian Kuehn. "Adaptive voter model on simplicial complexes." *Physical Review E* 101, no. 2 (2020): 022305.

[9] Calcaterra, Craig, and Wulf Kaal. *Decentralization: Technology's impact on organizational and societal structure*. Walter de Gruyter GmbH & Co KG, 2021.

[10] Calcaterra, Craig, Wulf A. Kaal, and Vlad Andrei. "Blockchain infrastructure for measuring domain specific reputation in autonomous decentralized and anonymous systems." *U of St. Thomas (Minnesota) Legal Studies Research Paper* 18-11 (2018).

[11] Calcaterra, Craig. "On-chain governance of decentralized autonomous organizations: Blockchain organization using Semada." *Available at SSRN 3188374* (2018).

[12] gitlab.com/dao-governance-framework/

[13] https://daogovernanceframework.com

[14] https://github.com/Lakat-OS/lakat-py