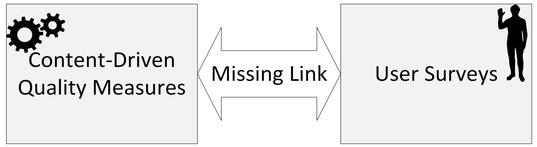
A Vision for performing Social and Economic Data Analysis using Wikipedia's edit history

Erik Dahm, Moritz Schubotz, Norman Meuschke, Bela Gipp
Dept. of Computer and Information Science, University of Konstanz, Universitätsstr. 10, 78457 Konstanz, Germany

www.isg.uni.kn



Introduction

There are **two distinct lines of research** regarding Wikipedia that are currently independent of each other. The first line of research includes content-based approaches that analyze Wikipedia's edit history to assess **contributor reputation and content quality**. The second line of research comprises user surveys studying **contributor motivation**, **contributor interaction**, and other factors that influence the quality of contributions to Wikipedia.

We suggest that analyzing Wikipedia's edit history and **linking** this data to individual characteristics of contributors collected through surveys could provide a large-scale, open source dataset offering tremendous potential for user-centered and content-centered research.

User-centered Studies

Contributor Motivation
The top motives for editing Wikipedia are "fun" and "ideology" with all other motives being significantly weaker in comparison [23].
Anthony et al. [5] show that registered and anonymous users differ in the number of edits they perform, the contribution size and the retention rate. This difference is assumed to be due to different motives (strong commitment to community and building up reputation).

Network Characteristics
Brandes et al. [8] showed that structural measures derived from the interactions between contributors (deletions, undeletions, and restorations) and the roles that the contributors play in the editing and revision process can be associated with article quality.

Editor Types
Liu and Ram [20] categorized contributors based on actions the perform (e.g., insertions, modifications, or deletions), where they were able to associate the composition of contributor types working on an article with the article's quality.
Yang et al. [30] extend the previous mentioned approach and also consider indirect work in non-article namespaces, like discussions about changes on articles or direct communication between contributors [30]. They find eight contributor roles that had a significant effect on the prediction of article quality. They also fount that articles can profit from different contributor roles in dierent stages of an article's life cycle.

Integrated Analysis of the Editorial Process in Wikipedia

We suggest that linking content-based analyses and user centered surveys can provide a dataset that offers two major benefits. First, the linked dataset enables the investigation of new research questions from domains such as the social sciences, economics and business, and computer science. Second, the large, information-rich, yet freely and openly available dataset would enable stakeholders, who could otherwise not obtain a comparable dataset, to perform big data analytics.

User Modeling
The Wikipedia edit history could provide insights on the popularity of articles, i.e. topics, and its development over time. Linking this data to user accounts and their demographics additionally allows to deduce the interests of specific users and creating user group profiles, i.e. topics that are likely relevant for specific user groups.

Team Composition
Comparing the findings from the online collaboration scenario in Wikipedia to observations of real-world collaboration experiments could yield valuable new insights for team composition and team efficiency research. It would be particularly interesting to see whether the contributor type determined from analyzing Wikipedia's edit history allows to predict and improve real-world team behavior, effectiveness and efficiency.

Collective Behavior
The approaches to investigate user modeling and team composition described above can be generalized and serve to investigate the collective behavior of Wikipedia contributors in general.
For example, the formation process of rules and guidelines is hard to fully explain by analyzing the users' edit history alone [9, 18]. Linking user survey data could uncover hidden variables that might explain certain phenomena of the regulation.

Reputation and Quality Analysis in Wikipedia
Establishing the "missing link" between content-based measures of reputation and article quality and the characteristics of the individuals behind the Wikipedia user accounts could enable investigations of the characteristics of successful Wikipedia contributors.

Expert Search
Existing user surveys show that many domain experts actively contribute to Wikipedia. Predicting domain expertise using content-based reputation measures and conforming this prediction in a targeted user survey could yield a valuable expert search system.

Talent Scouting
The idea of using Wikipedia editing data to find domain experts can be extended to the problem of talent identification. Especially domains suffering from a shortage of skilled personnel need effective and efficient means to identify future talents as early as possible. Linking user profiles with high content-based performance scores to the current demographic and socioeconomic properties of the individuals may enable the development of models suitable to predict future talents.

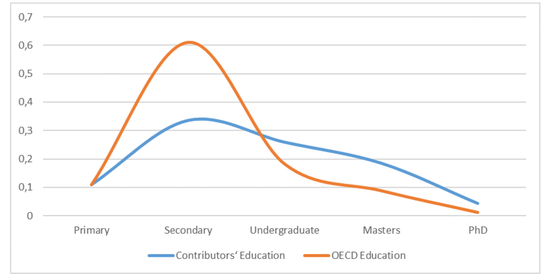
Content-based Assessments

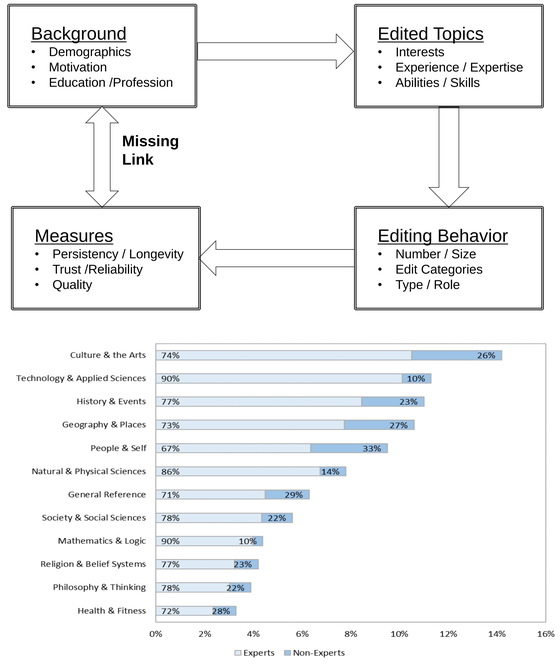
A simple quantification for the amount of contributions is the **edit count** (https://en.wikipedia.org/wiki/Wikipedia:Edit count) that Wikipedia provides for every user. However, the edit count considers each contribution with equal cardinality. Therefore, the informative value of the mere edit count for assessing contributors' reputation is very limited.

Productivity of authors
Wöhner et al. [29] compared different metrics for contributor reputations. The metric "efficiency of authors", which expresses the ratio of persistent contributions (that survive at least two weeks) to all contributions of a contributor was found to be best performing metric for measuring contributor reputation.

H-index
Suzuki [28] adapts the idea of the h-index [16] for assessing contributor impact in Wikipedia. Thereby, implicit positive ratings (contribution remains unaltered during revision) on the quality of a contribution are extracted from the edit history. Thus, amount of edits, the number of edited articles, and the quality of edits are considered.

WikiTrust
Here, the main idea of the Wikitrust [1] approach is to increase the reputation ratings of contributors if the edits they performed are preserved by subsequent contributors. Thereby, the "trustworthiness" of text is computed as a function of the reputation of the original contributor and the reputation of all contributors who edited the article in the proximity of the text [2]. Therefore, the WikiTrust reputation system allows automatically computing a contributor's reputation and estimating the trustworthiness of contributed text.





References

[1] B. T. Adler. WikiTrust: Content-driven Reputation for the Wikipedia. PhD thesis, UC Santa Cruz: Computer Science, 2012.
[2] B. T. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning Trust to Wikipedia Content. In Proc. WikiSym, pages 26:1{26:12, 2008.
[3] D. L. Anthony, S. W. Smith, and T. Williamson. Reputation and Reliability in Collective goods: The Case of the Online Encyclopedia Wikipedia. Rationality and Society, 21(3):283-306, 2009.
[4] J. Beel, S. Langer, A. Nürnberger, and M. Genzmehr. The Impact of Demographics (Age and Gender) and other User-Characteristics on Evaluating Recommender Systems. In Proc. TPDL, pages 396-400, 2013.
[5] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network Analysis of Collaboration Structure in Wikipedia. In Proc. WWW, pages 731-731, 2009.
[6] B. Butler, E. Joyce, and J. Pike. Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in Wikipedia. In Proc. SIGCHI, pages 1101-110, 2008.
[7] J. E. Hirsch. An index to quantify an individual's scientific research output. PNAS, 102(46), 2005.
[8] D. Jemielniak. The SAGE Handbook of Action Research, chapter Naturally Emerging Regulation and the Danger of Delegitimizing Conventional Leadership: Drawing on the Example of Wikipedia, pages 522-528. Sage, 2015.
[9] J. Liu and S. Ram. Who Does What: Collaboration Patterns in the Wikipedia and their Impact on Article Quality. ACM Trans. Manage. Inf. Syst., 2(2):11:1-11:23, 2011.
[10] O. Nov. What Motivates Wikipedians? Commun. ACM, 50(11):60-64, 2007.
[11] Y. Suzuki. Quality Assessment of Wikipedia Articles Using h-index. Information Processing, 23(1), 2015.
[12] T. Wöhner, S. Köhler, and R. Peters. Automatische Reputationsmessung in der Wikipedia. In . WI, 2011.
[13] D. Yang, A. Halfaker, R. Kraut, and E. Hovy. Who Did What: Editor Role Identication in Wikipedia. In Proc. Int. AAAI Conf. on Web and Social Media, 2016.
[14] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertesz. Dynamics of Conicts in Wikipedia. PLOS ONE, 7(6):1-12, 06 2012.