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
Universitaetsstr. 10, 78457 Konstanz, Germany
firstname.lastname@uni-konstanz.de

ABSTRACT

In this vision paper, we suggest combining two lines of research to study the collective behavior of Wikipedia contributors. The first line of research analyzes Wikipedia's edit history to quantify the quality of individual contributions and the resulting reputation of the contributor. The second line of research surveys Wikipedia contributors to gain insights, e.g., on their personal and professional background, socioeconomic status, or motives to contribute to Wikipedia. While both lines of research are valuable on their own, we argue that the combination of both approaches could yield insights that exceed the sum of the individual parts. Linking survey data to contributor reputation and content-based quality metrics could provide a large-scale, public domain data set to perform user modeling, i.e. deducing interest profiles of user groups. User profiles can, among other applications, help to improve recommender systems. The resulting dataset can also enable a better understanding and improved prediction of high quality Wikipedia content and successful Wikipedia contributors. Furthermore, the dataset can enable novel research approaches to investigate team composition and collective behavior as well as help to identify domain experts and young talents. We report on the status of implementing our large-scale, content-based analysis of the Wikipedia edit history using the big data processing framework Apache Flink. Additionally, we describe our plans to conduct a survey among Wikipedia contributors to enhance the content-based quality metrics.

Keywords

Wikipedia; Author Reputation; Article Quality; Editor Types

1. INTRODUCTION

Wikipedia is the largest collaboratively maintained information repository on the Web. The Wikipedia contains more than 40 million articles[1] and attracts billions of annual visitors[2]. Wikipedia's openness that allows virtually everyone to contribute and edit content is a key factor that ensures the breadth, diversity, and currentness of Wikipedia's content, which in turn is a driving force of Wikipedia's success. However, Wikipedia's open and collaborative editing process is also a source of doubt regarding the quality and reliability of Wikipedia content

Assessing the reputation, i.e. "quality", of Wikipedia contributors and the quality of Wikipedia content are problems that have attracted much research attention in recent years. Having reviewed the literature on approaches that analyze the editing and revision process of Wikipedia (see Section 2), we found two distinct lines of research that are currently independent of each other (Figure 1). The first line of research includes content-based approaches that analyze Wikipedia's edit history to assess contributor reputation and content quality (see Section 2.1). The edit history represents a persistent, fine-grained record of any change to an article and the originator of the change. Content-based approaches analyze Wikipedia's edit history to assess or predict the trustworthiness of contributors, the quality of their contributions, and the overall quality of Wikipedia articles. All content-based investigations of quality issues in Wikipedia that we found rely on IP addresses or user account names to distinguish individual contributors. These investigations allow for little conclusions regarding the individuals the accounts represent. Content-based analysis approaches using Wikipedia's edit history yield valuable results for assessing and ensuring content quality in Wikipedia, yet do not allow linking this data to individuals.

The second line of research comprises user surveys studying contributor motivation, contributor interaction, and other factors that influence the quality of contributions to Wikipedia (see Section 2.2). While some surveys investigated socioeconomic questions in regard to Wikipedia users, this data is not linked to accounts or IP addresses, which would allow to model the behavior of the individuals.

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. The data set would enable to investigate questions such as:

- How do user characteristics, e.g., demographics, influence the relevance of topics to the user?

[1] https://en.wikipedia.org/wiki/Wikipedia:Statistics

[2] Wikipedia currently holds rank 6 in the Alexa traffic ranking http://www.alexa.com/siteinfo/wikipedia.org

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3053363

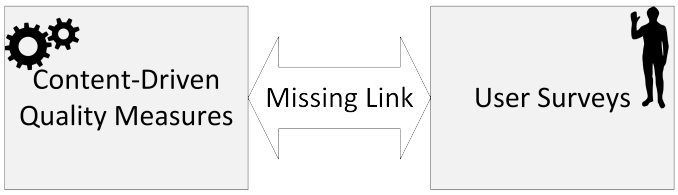

Figure 1: Existing research either analyses the quality of Wikipedia's content using automated procedures or investigates the behavior of Wikipedia users with the help of traditional surveys. Performing the types of social and economic data analysis we envision requires linking the two data sources.

- How does user-specific relevance develop over time?
- Can one derive user models that predict the relevance and relatedness of topics for users and user groups?
- How can user models improve information retrieval systems, such as content and item recommender systems [7]?
- How does the interaction of user accounts observable in the edit history relate to interaction patterns of individuals in real-world situations known from sociology [32]?
- Can one predict the career paths of young contributors based on the edits they perform?
- Can one identify domain experts by analyzing the Wikipedia edit history?
- Can one estimate socio-economic properties of individuals such as education, profession, or social status by analyzing Wikipedia's edit history?

To explain our vision of how these and other research questions could be answered, we structure the remainder of this paper as follows. In Section 2, we review existing research that investigates Wikipedia's edit process and Wikipedia contributors. In Section 3, we explain the potential benefits of linking data from Wikipedia's edit history and survey data to enable novel research approaches in several areas of the social sciences, business and economics, and computer science. In Section 4, we present the current status of our technical solution for analyzing Wikipedia's edit history and our efforts to perform tailored user surveys to complement and extend the insights derived from our automated content-based analysis. Section 5 concludes the paper with an outlook on future research.

2. RELATED WORK

Our objective is to link content-based quality assessments derived from analyzing the Wikipedia edit history to specific characteristics of Wikipedia contributors collected through surveys. Given this objective, this Section reviews related work from the two lines of research in the context of Wikipedia that are currently independent of each other, but shall be linked in our approach. In Section 2.1, we review content-based approaches to assess contributor reputation and content quality in Wikipedia. In Section 2.2, we review studies that investigated user-centered factors that influence contributor behavior and content quality. In Section 2.3, we briefly describe the design and results of the first global Wikipedia Survey [13], which we plan to take as an example for our own user survey.

2.1 Content-based Assessments of Contributor Reputation and Article Quality

Reputation is typically defined as the public opinion towards a person, a group of persons, or an organization derived from the social evaluation of a set of criteria [17]. The increasing amount and of user-generated content on the Web has increased the importance of establishing and quantifying reputation for Web users. Reputable users can be characterized as users who regularly provide high quality content that is useful for many other users. Reputable users are an essential asset for many Web sites, such as online forums, blogs, and wikis [10]. Being the largest collaborative information repository, determining user reputation is of special importance to Wikipedia. The task has attracted much research, which we briefly review in the following.

Key components of the approaches that have been proposed to measure contributor reputation in Wikipedia correspond to well-established factors used to quantify reputation in academia. Quantifying the productivity, quality, and impact of research contributions for researchers or research institutions is a well-established process. Academic quality metrics are important input data for numerous decision making processes, such as the hiring and promotion of researchers, the funding of research projects, or the ranking of research institutions. The most widely-used indicators of academic reputation consider bibliometric data, i.e. data on the published research works and the number of citations these works have received. Bibliometric data is at the heart of indicators quantifying the reputation of individual researchers or research institutions, such as the h-index [16]. This index assigns a high value to researchers or institutions who publish many research works that are highly cited by other researchers. Bibliometric data is also the base for computing indicators to quantify the reputation of academic venues, such as the impact factor [12]. Several researchers question the informative value of such measures [22, 27] as well as the transparency [21] and fraud-resilience of their computation [6]. However, thus far, no better approach for quantifying reputation and productivity in academia has found wide-spread use.

Some use cases allow transferring bibliometric approaches to Wikipedia by equating the concepts 'publication', 'author', and 'citation' with the corresponding concepts 'article', 'contributor', and 'intra-wiki-link'. However, for the use case of reputation analysis, bibliometric indicators can only partially be transferred to Wikipedia for several reasons.

First, Wikipedia contributors typically do not author entire articles. Rather, Wikipedia articles evolve through a continuous collaborative editing and revision process that typically involves several contributors. Therefore, in most cases, only parts of an article can be attributed to a specific contributor. The authorship attribution problem is a well-known challenge for Wikipedia research (see e.g. [11]) and complicates the transfer of bibliometric concepts like considering a contributor's number of articles.

Second, the concept of casting a quality judgment by citing another work, which is essential to Bibliometrics, is not directly transferable to Wikipedia. In Wikipedia, topical relatedness, not article quality is the dominant reason for cross-referencing other articles. Therefore, links in Wikipedia cannot assume the role that academic citations have for assessing the reputation of contributors or the impact of articles. In Wikipedia, the number of article views can

be seen as an impact measure for an article. However, in many cases it is debatable whether article views are more indicative of article quality and impact or rather indicators of topical popularity [26].
Third, the quality of Wikipedia articles typically cannot be judged immediately at the time of article creation. Most academic publications undergo an editorial or peer review process, which ensures the quality of the publication as a whole prior to publication. Wikipedia articles continuously develop over time and many start out as primitive stubs. Therefore, the quality of Wikipedia articles and edits is assured through a series of continuous procedures, such as collaborative review of edits or limiting editing rights to selected users, rather than through a quality check of the entire article at fixed points in time. Due to this characteristic of the Wikipedia editing process, many quality metrics for Wikipedia articles consider the longevity of (unchanged) contributions as part of the revision process.
Although classic bibliometric measures cannot directly be applied in the context of Wikipedia, many contributor reputation metrics for Wikipedia also reflect the notion of productivity and quality that is at the heart of bibliometric measures. In other words, the more content someone contributes to Wikipedia, and the higher the quality of this content, the higher will be the reputation assigned to the contributor. Sections 2.1.1 to 2.1.3 describe content-based approaches that consider the outlined characteristics to assess the quality of contributions, and by doing so, the reputation of contributors as well as the overall article quality.

2.1.1 Productivity of Contributors

As we describe in the previous Section, approaches to determine contributors' reputation typically aim at quantifying the amount and the quality of contributions. While quality seems to be a subjective characteristic (at least at first) and therefore hard to quantify, the frequency and amount of contributions is easy to assess. A simple quantification for the amount of contributions is the edit count[3] that Wikipedia provides for every user. This number reflects all changes a user submitted to Wikipedia and is typically seen as a measure of a user's experience within the Wikipedia community. However, the edit count considers each contribution with equal cardinality, regardless of the contribution type, e.g., insertion, deletion, or revert, and other characteristics, e.g., length or longevity of the contribution. Therefore, the informative value of the mere edit count for assessing contributors' reputation is very limited.
Wöhner et al. [29] compared different content-based metrics for measuring contributor reputation found in the literature, e.g., the number of edits and contributions to high quality articles with several newly-developed metrics, e.g., the number of persistent words, the size of the largest persistent contribution, and the share of persistent contributions. They compared the metrics in regard to their discriminatory power for classifying "good" and "bad" contributors, i.e. contributors that were blocked due to violations of Wikipedia's policies. They found that the efficiency of contributors is the best performing metric for this task. The metric expresses the ratio of persistent contributions, i.e. contributions that "survive" at least two weeks, to all contributions of a contributor. Thereby, this metric explicitly includes both the frequency and quality of contributions.

[3]https://en.wikipedia.org/wiki/Wikipedia:Edit_count

2.1.2 H-Index with P-Ratio

Suzuki [28] adapted the h-index [16] to assess contributor impact in Wikipedia. The classic h-index known in academia quantifies the number of publications h that received at least h citations from other publications. Therefore, the index reflects an author's publications with the most impact.
Suzuki used this idea for assessing text, contributor, and article quality in Wikipedia. The approach extracts implicit positive ratings on the quality of a contribution from the edit history. If a contribution remains unaltered during review, this is considered as an implicit positive rating. He uses these implicit ratings to define the adapted h-index of a contributor as follows: $h(e)$ is the index of a contributor e, who edits more than $h(e)$ articles within which he or she receives positive ratings from at least $h(e)$ contributors. In the other $(N - h(e))$ articles, e receives non-positive ratings from less than $h(e)$ contributors. The approach thus considers the amount of edits, the number of edited articles, and the quality of edits. Additionally, he introduces a ratio of the h-index and the number of edited articles (p-ratio) to distinguish between vandals and low frequency contributors. He compares the h-index metric to manual ratings of Wikipedia contributors and found a strong correlation.

2.1.3 WikiTrust

WikiTrust [1] is another noteworthy content-based metric to assess contributor reputation. As in the case of the adapted h-index, the main idea of the WikiTrust approach is to increase the reputation ratings of contributors if the edits they performed are preserved by subsequent contributors. Respectively, contributors lose reputation when their edits are rolled back or are undone soon after the edit [3]. The longer text stays unchanged by other contributors (edit longevity), the higher is the assumed quality and trustworthiness of the text. 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].
WikiTrust distinguishes between contributing text to articles and editing contributions of others [1]. To determine the number of contributions and edits, the edit history of an article is examined. Contributions are considered useful if they survive multiple revisions. Additionally, the size of the contribution and the reputation of the revising contributors are taken into account. Longer contributions, as well as contributions that receive implicit approval by high quality contributors, i.e. remain unaltered, have a higher impact on the contributors' reputation. The reputation ratings for contributors are iteratively updated for each revision and each version of an article to reflect how long the contributions and edits remained unaltered in successive versions.
In an empirical analysis, *Adler et al.* found that short-lived text contributions and edits strongly correlate with low-reputation contributors while higher reputation scores of contributors correlate with a longer expected life-span of words [2]. More precisely, for contributions by low-reputation contributors, as judged by the reputation system, the likelihood of being short-lived is four times as high as the average likelihood of being short-lived [1, p. 151].
Therefore, the WikiTrust reputation system allows automatically computing a contributor's reputation and estimating the trustworthiness of contributed text.

2.2 User-centered Studies

In this subsection, we review studies that investigated the motives of contributors, the impact of motivational factors on the behavior of the contributors and the contribution quality, the influence of the network structure of collaborating contributors on article quality, and the classification of contributor types based on their edit behavior.

2.2.1 Contributor Motivation

An important question for research on Wikipedia is why contributors become and stay active and how different motives for contributing relate to editing behavior. *Oded Nov* investigated motivational reasons for editing Wikipedia and related them to high or low levels of contributing [23]. He found that the top motives for editing Wikipedia are "fun" and "ideology" with all other motives being significantly weaker in comparison. Additionally, the author finds that all motivational categories are positively correlated to the weekly hours that contributors invest.

Yang and Lai [31] improved on *Oded Nov's* study by constructing an integrated motivation model to determine the most important motivational factor. Factors they assume to positively influence knowledge sharing behavior are intrinsic motivation, extrinsic motivation, external self-concept, and internal self-concept. A structural equation model reveals that only internal self-concept is found to be relevant.

In a later study, *Yang and Lai* extended their model by employing expectation-confirmation theory and expectancy-value theory [19]. They found a positive relationship between confirmation, subjective task value, and individual satisfaction with contributing.

Anthony et al. [5] related motives of Wikipedia contributors to the types of contributions made. They compared registered and anonymous users assuming that registered users exhibit a strong commitment to the community and are mainly motivated by building up reputation in the community. Their results show that both contributor groups differ in the number of edits they perform, the contribution size and the retention rate. The contributions of registered contributors are significantly more frequent and larger in size than the contributions of anonymous contributors. Nevertheless, anonymous users with fewer contributions have a higher reliability compared to both anonymous and registered users with more contributions.

In summary, the presented studies show that contributors with higher intrinsic motivation spend significantly more time editing Wikipedia content. Furthermore, motivational differences significantly affect the reliability of contributions.

2.2.2 Network Analysis

Brandes et al. [8] analyzed collaboration among Wikipedia contributors from a network analytical perspective. They established a network of Wikipedia contributors working on an article, in which nodes represent the contributors and edges represent their interactions.

Three types of interactions between contributors are considered for establishing the network: deletions, undeletions, and restorations. The authors applied network analytical measures and showed that structural network parameters are correlated with article quality labels assigned by Wikipedia contributors, such as featured or controversial articles. They showed that structural measures derived from the interactions between contributors and the roles that the contributors play in the editing and revision process can be associated with article quality. While some contributors focus on providing content, others focus on reviewing edits. Each contributor type fulfills valuable functions within the editing and revision process.

2.2.3 Contributor Types

Liu and Ram [20] empirically explored the relationship between the collaboration of different contributor types and the resulting article quality. They categorized contributors based on the actions they perform, e.g., insertions, modifications, or deletions, and associated the composition of contributor types working on an article with the article's quality. They found six clusters of different sizes, which they assigned to role labels representing the predominant actions performed by the cluster: i) *All-round Contributors* are engaged in almost all types of actions, ii) *Watchdogs* perform mostly reverts, iii) *Starters* mostly create sentences consisting only of plain text, iv) *Content Justifiers* mainly add links and references, v) *Copy Editors* mainly modify existing sentences and vi) *Cleaners* mainly remove incorrect sentences, links, and references. Contributors can assume a different role for different articles.

Clustering articles by the types of the collaborating contributors yielded five clusters. Assessing the quality of the articles in those clusters showed that collaboration patterns and article quality, as judged by the Wikipedia community, are strongly correlated. Another analysis showed that the collaboration pattern has a significant impact on article quality, even when controlling for confounding variables.

Yang et al. [30] extended the idea of *Liu and Ram* by relating contributors' roles and their collaboration patterns to contribution and article quality. They first defined a set of social roles that contributors can assume using edit categories. Their taxonomy for categorizing edits is built upon actions, e.g., insert, delete, or modify. New in their approach is that they also consider indirect work in non-article namespaces, like discussions about changes on articles or direct communication between contributors [30].

Yang et al. found eight contributor roles that had a significant effect on the prediction of article quality. They also found that articles can profit from different contributor roles in different stages of an article's life cycle. This supports the assumption that contributors fulfill different functions.

Both papers show that different types of contributors and different collaboration patterns are identifiable in Wikipedia's edit history. The papers also show the existence of a connection between collaboration patterns and quality. This supports the importance of teamwork among contributors, as well as the importance of different types or roles fulfilling specific functions during the editing process.

Analyzing contributor types requires accounting for the influence of bot edits. This task is not trivial, since there is a continuous spectrum between human and bot contributions that complicates effective bot identification. For instance, in manual investigations, we noticed that the user with the pseudonym *Ron Meier*[4] improved the literature references for the article on the *Nakajima-Zwanzig equation* related to quantum mechanics in the German Wikipedia. For this purpose, Ron Meier employed a large set of regular expressions that he actively maintains and also uses to edit other articles. On Jan. 31, 2017, he performed 28 edits throughout

[4]https://de.wikipedia.org/wiki/Benutzer:RonMeier

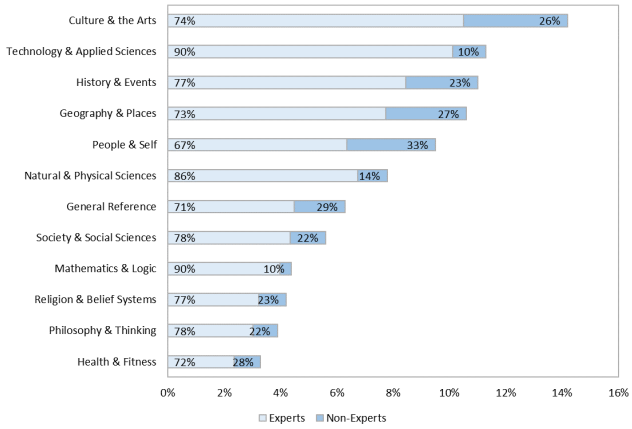

Figure 2: Results of the global Wikipedia user survey: distribution of edited fields and the fraction of experts that edit the particular fields.

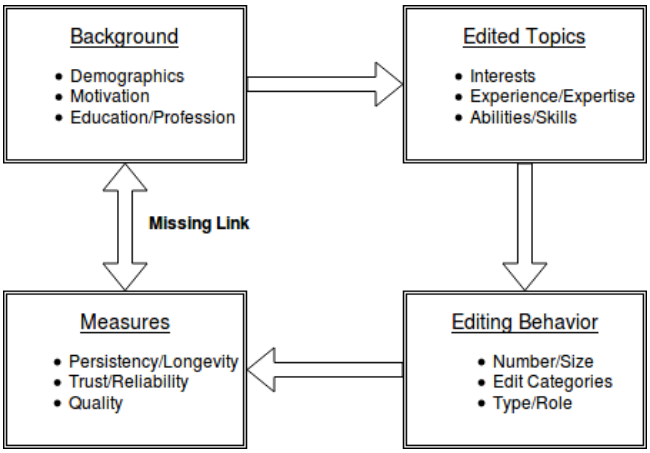

Figure 3: Connection between the background of the contributors, the edited topics, the editing behavior, and quality measures.

the day. Thereof, 16 edits were marked as minor and 14 edits concerned the formatting of literature references. Ron Meier also edited 3 discussion pages to reflect that he had fixed a broken Web-link, which was reported by the so called giftbot. The aforementioned characteristics indicate that Ron Meier falls into the contributor category *Cleaners*. However, some automated approaches might misclassify Ron Meier's edits as bot edits due to the high editing speed achieved through the use of regular expressions.

2.3 Global Wikipedia Survey

The characteristics of Wikipedia contributors, have been investigated in several surveys, including the semi-annual editor survey conducted by the Wikimedia Foundation[5]. This regular Wikimedia survey covers contributors' demographics, motives for editing, and editing activity as well as the overall spirit within the community. Some of the surveys also addressed specific issues such as the welcoming culture among contributors[6]. The first global survey of Wikipedia contributors was conducted by researchers of the UNU-MERIT research institute in 2009 [13]. The survey asked for the demographics, activity, and motives of contributors. Additionally, contributors were asked to rate their expertise in the topics to which they contributed. We briefly summarize the findings of this survey, as they are particularly relevant for our approach.

The survey received 176.192 responses. Almost 31% of respondents contribute actively to Wikipedia (23.25% are occasional contributors and 7.42% are regular contributors). However, the majority of 65.92% are readers. The average contributor is 26.1 years of age. The age distribution is highly skewed, with more than half of the respondents being younger than 22 years. Nearly 50% of the contributors have completed tertiary education, i.e. hold an undergraduate, master's, or PhD degree. Asked about their motivation to contribute to Wikipedia, most contributors selected "I like the idea of sharing knowledge and want to contribute to it" (72.91%) and "I saw an error I wanted to fix" (68.78%) as their top-ranked answers. The items that respondents agreed least with include: "To improve my job / career opportunities" (1.71) and "To gain a reputation in the Wikipedia community" (2.16%) [15]. In every field, except for "People & Self", the share of contributors claiming to have expertise in that field is above 70%. In the fields "Mathematics & Logic" (90.45%), "Technology & Applied Sciences" (89.54%) and "Natural & Physical Sciences" (86.45%), almost all contributors claim to have specific background knowledge in the respective field (see Figure 2).

Respondents were asked whether they had undergone formal training or acquired work experience in the fields to which they contributed. 48.8% stated to have formal training and 45.6% to have work experience [14]. Given these results, we hypothesize that linking data on the editing performance of Wikipedia contributors to their educational and professional background is of particular interest to better predict contribution quality.

3. INTEGRATED ANALYSIS OF THE EDITORIAL PROCESS IN WIKIPEDIA

Section 2 shows that content-based approaches that analyze Wikipedia's edit history achieve good results in measuring and predicting the quality of content and the reputation of contributors. Likewise, a number of user-centered studies, mostly in the form of surveys, provide valuable insights on the personal characteristics of contributors and on group-dynamics that influence content quality 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. For instance, small enterprises or NGOs could use the linked dataset to perform market research or impact analyses. In the following, we describe a number of research

[5]https://meta.wikimedia.org/wiki/Research:Wikipedia_Editors_Survey_2011_April

[6]See for example: https://commons.wikimedia.org/wiki/File:Analysis_Wikimedia_Germany_Editor_Survey_2016_on_Welcoming_Culture.pdf

areas and research questions that could be investigated using the linked dataset.

User Modeling Global players such as Google, Amazon, Ebay, Facebook, or Twitter have access to large datasets of user transactions that allow for fine-grained analyses of user interests, user needs, and consumer behavior. Those proprietary datasets significantly contribute to improving the services of the respective companies, thus represent a major economic advantage over small and medium size enterprises. The Wikipedia edit history can partially alleviate this difference by providing 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.

The linked dataset could enable user modeling to improve a variety of services. For example, performing user modeling using this public domain data could help to overcome the *cold start problem* in recommender systems, i.e. the lack of sufficient transaction data to generate useful recommendations. Typically, items, e.g., in online shops, are recommended to users based on the behavior of "similar" users. This process, known as collaborative filtering, requires a certain amount of user transactions to yield good results. Linking user-specific data to the users' edit behavior recorded in Wikipedia's edit history could enable the deduction of non-trivial, user-specific relations between topics and items. The identified relations could reduce the cold start problem.

To illustrate the approach, we construct a fictive example for possible topic relations that might characterize specific user groups and could be identified from analyzing users' demographics in conjunction with the users' edit behavior in Wikipedia. We consider the characteristics age, gender, education, income, and country of residence. The analysis might show that male Wikipedia contributors, aged 45-60, who underwent tertiary education, reside in Norway and have an annual income above 150,000 USD more frequently edit articles on the PGA Golf Tour, luxury yacht models, and helicopter skiing providers than contributors from other groups. Female contributors residing in Norway, who are in the 15-20 age group, currently undergo secondary education, and have an annual income below 5,000 USD might more frequently than other groups contribute to the articles on the YWCA-YMCA Guides and Scouts of Norway, tourist attractions at the Spanish Costa Brava, and YouTube personalities hosting channels in the beauty and fashion segment.

Providers of recommender systems could use the knowledge about user group interests derived from Wikipedia. For example, female Norwegian teenagers might be recommended information, products and services related to camping equipment, outdoor apparel, cosmetics, fashion, packaged tours, and mobile devices, whereas middle-aged Norwegian males might be recommended yacht equipment, luxury cars, off-shore fishing tours, equipment for powder skiing, or popular golf resorts.

Team Composition Two recent articles [20, 30] suggest that the overall quality of Wikipedia articles is significantly higher if certain combinations of contributor types participated in the editing process of an article. 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. For example, would teams that include the various contributor types observable in Wikipedia also perform better in real-world scenarios? Can determining specific contributor types observable in Wikipedia facilitate the composition of real-world teams? These and other research questions could be investigated if content-based observations were linked to specific user accounts.

One example that suggests the findings from observing successful collaboration in Wikipedia could be transferable to offline collaboration scenarios is the edit history of the formula for the mean reciprocal rank. MRR is a well-known metric to evaluate known item information retrieval tasks. The formula for calculating MRR (1) was added to the English version of Wikipedia on August 22, 2008 by Elif Aktolga[7], who was a PhD student in computer science at the University of Massachusetts, Amherst at that time:

$$MRR = \frac{1}{|Q|} \sum_{i=1}^{Q} \frac{1}{rank_i} \tag{1}$$

The original version of the formula (1) violated the notation convention established in mathematics to typeset multi-character operators, such as MRR in upright font. This deficit was corrected by the mathematician Michael Hardy[8] on March 15, 2009 resulting in the current version of the formula (2).

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}. \tag{2}$$

The example demonstrates how the cooperation of individuals from different backgrounds can improve the quality of the final result at a high level of detail. We hypothesize that stimulating similar successful collaboration is possible in real-world situations.

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 that describe how users are supposed to edit Wikipedia 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 process, such as spontaneous rule enactment [18].

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. This would

[7] https://en.wikipedia.org/w/index.php?title=Mean_reciprocal_rank&oldid=233566401

[8] https://en.wikipedia.org/w/index.php?oldid=277355460

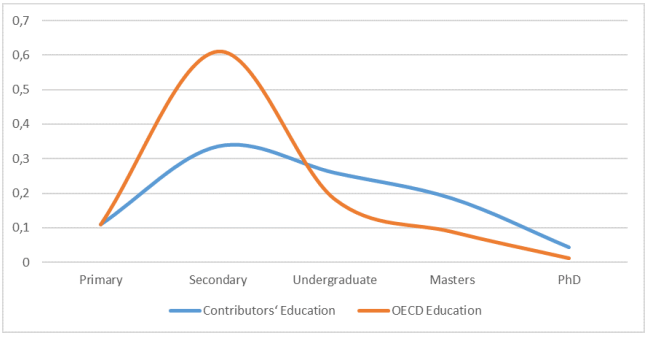

Figure 4: Comparison of the distribution of educational degrees among Wikipedia contributors (2010) and the OECD average (2010), according to the ISCED-97 classification.

allow for a better understanding of the emergence of user reputation and content quality and improve their prediction (c.f. Figure 3).

Expert Search Existing user surveys (see Section 2.3) 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. Especially for specialized areas of expertise, the availability of an algorithmic approach to retrieve promising candidates from a large and broad collection like Wikipedia could significantly reduce the effort for determining experts.

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.

4. IMPLEMENTATION

Existing research on content-based quality metrics (see Section 2.1) provides detailed information on the mathematical foundations of the metrics, yet little implementation details and no open source implementation to reproduce the results. Therefore, we re-implemented the methods presented in 2.1 and added specialized content metrics derived from the analysis of mathematical formulae and named entities in Wikipedia [24]. Our implementation uses the big data processing platform *Apache Flink* [4] and the HDFS file system. For the efficient analysis of formulae, we use Wikimedia's formulae rendering service *mathoid* [25]. Using big data processing technology is essential given the immense size of the dataset. For instance, the dump of the English Wikipedia that includes edit history data has a size of approx. 10 TB. To reduce memory consumption and computational load, we work on the compressed version of the dumps and use edit scripts as internal data model. We will soon publish a more detailed description of our framework.
A survey of Wikipedia contributors is currently in the planning stage. We will draw a reasonably sized sample from the population of registered Wikipedia contributors and invite them to a web-based survey. The questionnaire will collect account name and demographics, including nationality, sex, age, and language skills. The focus of the survey will be on the contributors' educational and professional background, their editing behavior and motives for editing. As described in Section 3, we are especially interested in connecting quantitative reputation measures with a contributor's professional and educational background. We hope to answer the question if and how a contributor's motivation and educational or professional background influence the set of topics edited. Furthermore, we will investigate potential differences in the quality of contributions to various topics. The questionnaire will follow the structure established in the global Wikipedia survey [15].
We are currently developing a statistical framework to compensate for the bias caused by the fact that Wikipedia editors are not a representative sample of society at large. Our goal is to develop adaptive sampling methods that enable compensating for this bias. Once we have trained our statistical model, we plan to reproduce well-known socioeconomic statistics, such as the OECD education distribution (Figure 4) from the distribution of the respective characteristics among Wikipedia contributors. In other words, we seek to approximate, e.g., the red curve in Figure 4, from the blue curve using our bias compensating sampling methods.
The planned user survey naturally raises questions about protecting the privacy of contributors. To guarantee privacy, we will provide survey participants absolute control over their data. We will enforce that no connection can be drawn between the account names and the data obtained from the survey. We will not publish account names with their corresponding reputation measures as this could potentially allow linking the accounts to the survey data. Nevertheless, in rare cases some contributors might be identifiable, e.g., through rare combinations of nationality, gender, educational background and edit count. To mitigate this risk, we will ensure that we and any other third party that might use the data will only report aggregated results.

5. CONCLUSION AND OUTLOOK

We propose linking content-based measures of contributor reputation and article quality in Wikipedia to contributor-specific data collected through surveys to enable the investigation of research questions in the social sciences, business and economics, and computer science. We motivate that linking the two data sources can enable novel research on user modeling, composing teams, analyzing collective behavior, searching for domain experts, and identifying potential talents early. We outline our implementation of content-based reputation and article quality metrics using the big data processing framework Apache Flink and briefly describe the planned web-based user survey.
Conducting surveys and other user-centered experiments are essential research methods to verify hypotheses in many fields. However, deriving statistically significant results from such studies requires large enough sample sizes. The temporal, organizational, and the financial effort required to conduct user experiments of sufficient size often poses significant challenges to researchers.
We envision that the user-centered analysis of editing behavior in Wikipedia might one day serve as a proxy for

large-scale behavioral observations of human subjects. Automated analyses of Wikipedia's edit history might serve as a large-scale "pre-check" of research hypotheses. Gathering evidence to support hypotheses with the help of cost-efficient big data processing technologies could initially serve as a preparatory step to identify hypotheses that are promising enough to conduct user experiments and surveys. To test this hypothesis, we are currently collaborating with researchers at the social science and economics department at the University of Konstanz to identify specific use cases to demonstrate the approach in practice.

6. 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] B. T. Adler, Thomas, and L. de Alfaro. A Content-driven Reputation System for the Wikipedia. In *Proc. WWW*, pages 261–270, 2007.

[4] A. Alexandrov, R. Bergmann, S. Ewen, J. Freytag, F. Hueske, A. Heise, O. Kao, M. Leich, U. Leser, V. Markl, F. Naumann, M. Peters, A. Rheinländer, M. J. Sax, S. Schelter, M. Höger, K. Tzoumas, and D. Warneke. The Stratosphere Platform for Big Data Analytics. *VLDB*, 23(6):939–964, 2014.

[5] 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.

[6] J. Beel and B. Gipp. On the Robustness of Google Scholar Against Spam. In *Proc. HT*, pages 297–298, 2010.

[7] 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.

[8] U. Brandes, P. Kenis, J. Lerner, and D. van Raaij. Network Analysis of Collaboration Structure in Wikipedia. In *Proc. WWW*, pages 731–731, 2009.

[9] 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 1101–1110, 2008.

[10] B.-C. Chen, J. Guo, B. Tseng, and J. Yang. User Reputation in a Comment Rating Environment. In *Proc. KDD*, pages 159–167, 2011.

[11] F. Flöck and M. Acosta. Wikiwho: Precise and Efficient Attribution of Authorship of Revisioned Content. In *Proc. WWW*, 2014.

[12] E. Garfield. Citation Indexes for Science: a New Dimension in Documentation through Association of Ideas. *Science*, 122(3159):108–111, 1955.

[13] R. Glott, P. Schmidt, and R. Ghosh. Wikipedia Survey – First Results. 2009.

[14] R. Glott, P. Schmidt, and R. Ghosh. Analysis of Wikipedia Survey Data. 2010.

[15] R. Glott, P. Schmidt, and R. Ghosh. Wikipedia Survey – Overview of Results. 2010.

[16] J. E. Hirsch. An index to quantify an individual's scientific research output. *PNAS*, 102(46), 2005.

[17] S. Javanmardi, C. Lopes, and P. Baldi. Modeling User Reputation in Wikis. *Statistical Analysis and Data Mining*, 3(2):126–139, 2010.

[18] 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.

[19] C.-Y. Lai and H.-L. Yang. The reasons why people continue editing wikipedia content - task value confirmation perspective. *Behaviour & Information Technology*, 33(12):1371–1382, 2014.

[20] 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.

[21] P. Meschenmoser, N. Meuschke, M. Hotz, and B. Gipp. Scraping Scientific Web Repositories: Challenges and Solutions for Automated Content Extraction. *D-Lib Magazine*, 22(9/10), 2016.

[22] H. Moed, W. Burger, J. Frankfort, and A. Van Raan. The application of bibliometric indicators: Important field- and time-dependent factors to be considered. *Scientometrics*, 8(3-4):177–203, 1985.

[23] O. Nov. What Motivates Wikipedians? *Commun. ACM*, 50(11):60–64, 2007.

[24] M. Schubotz, A. Grigorev, M. Leich, H. S. Cohl, N. Meuschke, B. Gipp, A. S. Youssef, and V. Markl. Semantification of Identifiers in Mathematics for Better Math Information Retrieval. In *Proc. SIGIR*, pages 135–144, 2016.

[25] M. Schubotz and G. Wicke. Mathoid: Robust, Scalable, Fast and Accessible Math Rendering for Wikipedia. In *Proc. CICM*, volume 8543 of *LNCS*, pages 224–235. 2014.

[26] M. Schwarzer, M. Schubotz, N. Meuschke, C. Breitinger, V. Markl, and B. Gipp. Evaluating Link-based Recommendations for Wikipedia. In *Proc. JCDL*, pages 191–200, 2016.

[27] P. O. Seglen. Why the Impact Factor of Journals Should Not Be Used for Evaluating Research. *BMJ*, 314(7079):497, 1997.

[28] Y. Suzuki. Quality Assessment of Wikipedia Articles Using h-index. *Information Processing*, 23(1), 2015.

[29] T. Wöhner, S. Köhler, and R. Peters. Automatische Reputationsmessung in der Wikipedia. In *Proc. WI*, 2011.

[30] D. Yang, A. Halfaker, R. Kraut, and E. Hovy. Who Did What: Editor Role Identification in Wikipedia. In *Proc. Int. AAAI Conf. on Web and Social Media*, 2016.

[31] H.-L. Yang and C.-Y. Lai. Motivations of Wikipedia content contributors. *Computers in Human Behavior*, 26(6):1377 – 1383, 2010.

[32] T. Yasseri, R. Sumi, A. Rung, A. Kornai, and J. Kertész. Dynamics of Conflicts in Wikipedia. *PLOS ONE*, 7(6):1–12, 06 2012.