Whose article is it anyway? – Detecting authorship distribution in Wikipedia articles over time with WIKIGINI

Fabian Flöck
Institute AIFB
Karlsruhe Institute of Technology
Karlsruhe, Germany
fabian.floeck@kit.edu

Andriy Rodchenko
Institute AIFB
Karlsruhe Institute of Technology
Karlsruhe, Germany
andriy.rodchenko@googlemail.com

ABSTRACT

In this work, we present a novel approach to detecting authorship of words in Wikipedia, which outperforms the baseline method in terms of accuracy. This is achieved by reducing the necessary word-based text-to-text comparisons, which are the most fallible steps in the process. To provide an aggregated measure of the concentration, we calculate a gini coefficient for each revision of an article based on our word-author-assignments. As a motivation for calculating this measure we argue that the concentration of words to just a few authors can be an indicator for a lack of quality and neutrality in an article. The coefficient development over time in an article is visualized and provided online as an easily accessible and useful tool to investigate how the content of an article evolved. We present examples where the gini curve gives useful insights into differences of articles and may help to spot crucial events in the past evolution of an article.

1. INTRODUCTION

A Wikipedia article is usually written by multiple editors. The composition of authors, their personal opinions and knowledge and the amount of text they contribute can vary tremendously, crucially influencing the quality and neutrality of the resulting article. Who wrote which content and how long did it last in the article? Who removed it and who reintroduced it? This information can give tremendous insights into how an article evolved. Its careful analysis can reveal how the content of an article – which most Wikipedia readers today implicitly assume to be vetted by the crowd and, hence, to be correct – actually reached its current state. It can reveal, for example, which specific parts where introduced by one of the first editors of the article and never have been changed notably. It can show if most of the editors of the article – apart from vandals – actually contribute to an article and help build its content; or if this is done by just a handful of authors. In this paper we discuss which possible loss of quality this authorship concentration might entail (Section 2) and propose a method to visualize and detect it. Therefore, we first revisit the current approaches to detecting word authorship in Wikipedia (Section 3.1) and propose a algorithm to do so more accurately (Section 3.2), as we show in our evaluation (Section 3.3). To create a meaningful aggregation and visualization of the distribution of word ownership in the article, we compute a "gini coefficient" as a measure of inequality (Section 4) and visualize its development over time (Section 5.1). We show some examples taken from the WIKIGINI tool to discuss its ability in helping detect anomalies in editing behavior (Section 5.2) and discuss further extensions of the tool as well as further planned research (Section 5.3).

2. WORD AUTHORSHIP DISTRIBUTION AS A QUALITY INDICATOR

The mechanism meant to secure the neutrality[1] and quality of an article is rooted in the checks-and-balances system of Wikipedia. It supposes that an article will (in most cases) have hundreds to thousands of users reading over it and correcting any possible biases and lack of quality: (i) The omission of relevant viewpoints, references or facts (including lack of timeliness),[2] (ii) the over-accentuation of certain points of view and facts that are not relevant enough for an representative overview of relevant information (required by Wikipedia), (iii) a polarized way of presenting specific viewpoints or facts (mostly through language), or (iv), on a meta-level, the existence or absence of whole articles about a topic.

Some articles in Wikipedia are built up and maintained by only a handful of people while others represent a more mass collaborative approach with tens to hundreds of users contributing each day. To achieve quality and neutrality, the task is to include all relevant facts, sources and viewpoints and write in a neutral tone, creating a balanced, representative article. We argue that in an article in which most of the words are written by only a very small percentage of all editors of that article it is – on average – much less likely that this task is accomplished than in the same article if it is authored by many different editors to the same extend. This is due to the fact that for just a few editors to achieve neutrality and balance in the content, it is necessary for these few to transcend their personal point of view and include all corresponding, relevant facts and sources in a timely manner. There are several obstacles to this: Single editors would have to have a comprehensive overview of all the existing relevant viewpoints, factual data and references on a specific topic and additionally would have to keep them up to date constantly. This might be possible in some highly specialized fields where only a few people publish data or voice their (relevant) opinions, or topics where it is feasible to collect *all* relevant facts and viewpoints through easily findable publications and other sources constantly. Yet, if this prerequisite is fulfilled, one would still have to assume that the main authors are completely free of bias of their own, i.e., they would not prefer one opinion or set of facts over the other (for example one opposed by themselves) when writing or updating the article. These assumptions outline the ideal picture of an editor and may hold for a fraction of all editors and edits present in Wikipedia, but are with a high probability to be rejected in most cases. To accomplish the discussed task, it is therefore much more likely that a multitude of – variously biased and imperfect – editors representatively covers all significant points of view in an article by collectively balancing out their biases, lacks in fact coverage and other quality impairments. In this light, it is reasonable to assume that a higher number of authors *actually contributing* to an article, on average, correlates with a lower risk of bias and incompleteness of the article, and vice versa. If, for example, 96% of all undeleted words in a specific article revision (with, say, 50 editors overall)

[1]Wikipedia's policy strongly promotes *Neutral Point Of View* in its articles, which *"means representing fairly, proportionately, and as far as possible without bias, all significant views that have been published by reliable sources"*. C.f. `http://en.wikipedia.org/w/index.php?title=Wikipedia:Neutral_point_of_view&oldid=477725395`.

[2]Note that the absence of certain facts or sources in an article can easily lead to a biased view on the topic, even if the presentation itself is neutral.

have been written by only 3% of its editors, it is much more likely to find lacking viewpoints, facts or sources in that article than in a revision where all 50 users have written around the same amount of words, as an equally distributed word authorship points to an active participation in the article by the editors instead of just lurking (or discussing) and leaving the main writing task to a "privileged" few. It also means that probably a lot of corrections have taken place as incremental steps to a better quality article (assuming the original authors were not without any fail).

In conclusion, the distribution of the authorship of written words in a specific article revision over the number of editors of that article can be a good indicator for the heightened risk of the article being incomplete and/or biased. This is, of course, only one among many possible indicators, some of which (and their interaction) we plan to examine in future work.[3]

Users who strive to actively engage in restoring neutrality and/or completeness to biased topics encounter the problem of how to determine if a selected article is biased if it is not already heavily discussed or appropriately tagged (which is often not the case for troubled articles).[4] Without the help of software tools, such users will have to go through the complete article discussion page and check all facts and phrasing in the text content to get a picture of the state the article is in. To have an easily understandable metric of which users contributed what portion of the article can make this task a lot easier. As it is, however, not straightforward to make sense out of the distribution of which users wrote how many words for all of the editors of an article, we implemented an aggregated overview showing a measure of inequality as a so called gini coefficient for each revision of an article in a timeline interface as we will describe in Section 5. This makes it easier to see at first glance if indicators of lack of quality exist. Additionally, it can serve as a more general monitor for detecting unusual editing patterns, going beyond the example cases and phenomena outlined in this section.

3. ASSIGNING WORDS TO THEIR AUTHORS

We determine which word in an analyzed article revision was written by which editor in what revision, appropriately handling insertion, deletion, rearrangement and reintroduction of text.

3.1 Related work

Several analysis and visualization tools have employed approaches to detect authorship information in Wikipedia article text. One of the earliest visualization tools was *HistoryFlow* by IBM, which assigns sentences of a text to the editor who created or changed them [6].[5] It provides a visual history stream of the authorship distribution by graphically representing for each revision the size of the text written by each author and also visualizing the changes from revision to revision. It doesn't however acknowledge deleted content, that was later reconstructed, as being written by the original editor. More importantly, by operating on a sentence level, small changes like spelling mistake corrections lead to wrongly recognizing the correcting editor as the author of the whole sentence.[6]

By tracking how long certain words remain unrevised in an article and which editors wrote those words, *Wikitrust* generates a visual mark-up of trusted and untrusted passages in any Wikipedia article text in different color shades [1, 2].[7]. To track authorship, longest matches for all word sequences of the current revision are searched for in the preceding revision and in previously existing, but now deleted word-chunks. In this way, Wikitrust can as well track reintroduced words and assign the original author. The underlying algorithm is, however, a variation of a greedy algorithm [1], known to look for local optima, which in the case of determining word authorship can lead to grave misinterpretations when word sequences are moved rather than inserted or deleted only [3]. Regarding computational cost, all undeleted and deleted word sequences have to be matched word by word to the most recent revision. Also, the tool is not intended to give an account of the distribution of words over editors in an article in the first place and therefore lacks any interface providing this exact information. This is however done by *WikiPraise*, a community extension using the results of the Wikitrust algorithm, extending the original functionality by displaying how much content belongs to which author. Its principal caveat is the kind of presentation of different word counts belonging to different editors and markups, given per revision. Users can hardly make sense of this copious, disparate, unlinked data, as it is complicated to interpret percentages of word authorship and their evolution over a range of revisions just by clicking through them.

3.2 Algorithm

To remedy most of the limitations of current approaches, we implemented a method determining authorship in Wikipedia (i) accurately, (ii) on a word level and (iii) considering reintroduction of formerly deleted text and rearrangements of text. The first assumption made for our approach is that most edits (if they are not vandalistic) change only a very minor part of an article's content. Hence, it is mostly unnecessary to compare the complete content of two article revisions with each other on a word level. Our algorithm thus splits up the article's wiki syntax into a tree structure, where all branches and leaves are assigned MD5 hash values. The article as a whole represents the stem, with each paragraph (separated by double line breaks) being a branch and all sentences in a paragraph denoting leaves. The words in each sentence are not assigned hash values. All hash values for the built article tree are stored in a hash table. When comparing an article with its preceding revision, the hash values for the stems, the branches and the leaves are checked for matches in the two article trees. In this way, paragraphs or sentences that have not been altered (even if they have been moved to another position in the text) are detected as the same (see Figure 1).[8] If a paragraph match has been found, checking of its sentences' hash values is skipped. If a matching hash value is not found for a paragraph, its sentences' hashes are checked for matches. Sentences and paragraphs matched via hash values inherit the authorship mark-up from the former revision. For those sentences with-

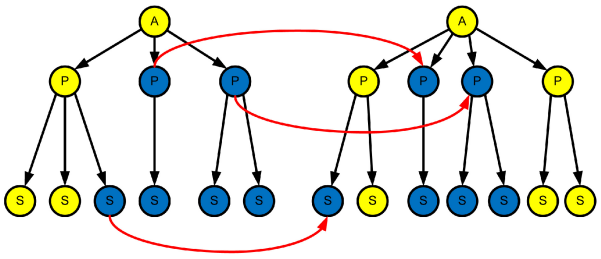

Figure 1: Two article trees of subsequent revisions (right: recent revision). Blue nodes represent unchanged paragraphs *P* or sentences *S*, yellow nodes are unique parts. Red arrows indicate inheritance of authorship.

[3]For other potential indicators see our work on bias-inducing sociotechnical mechanisms in Wikipedia [4].

[4]Only some potentially problematic articles get flagged with the respective warning templates by the Wikipedia community, like e.g. `http://en.wikipedia.org/wiki/Wikipedia:Neutrality_templates` Most templates are assigned to more popular articles, not covering the "long tail" of rather fringe topics.

[5]`http://www.research.ibm.com/visual/projects/history_flow/` No longer maintained.

[6]This is a grave imprecision issue, as a significant fraction of edits are devoted to grammatical and spelling mistakes correction.

[7]A "reputation system for Wikipedia authors and content", `http://www.wikitrust.net/`, via a browser extension.

[8]I.e., moving content to another position does not mean authoring that content, in our definition.

out MD5 hash matches, a vector is constructed from the contained words and their frequencies. In this subset, each sentence from the current revision is matched with the sentence in the preceding revision exhibiting the highest cosine similarity of vectors, with the similarity not being smaller than 0.71. In this way, sentences are detected which have been changed by a few words but essentially stayed the same. For sentences that cannot be matched in this way either, all the words in these sentences are marked as new (when only appearing in the current revision) or deleted (when from the preceding revision), respectively. Sentences matched by their vectors are compared using a standard text differentiation algorithm, namely the *difflib* library of Python, in which we implemented our approach. The outputs are inserted or deleted words in these sentences. After extensive testing, we discovered that the amalgamation of these techniques (article tree, vectors and text diff) yields the best combination of efficiency and accuracy.

Having determined the words added and deleted in a revision of the article, the editor of that revision is assigned as the owner of every word added in that revision that was not restored from a previous revision. This information (and the revision ID) is stored for all words in the article for every revision. All words unchanged or simply moved keep their original authorship mark-up. As the hash comparison not only looks for matches in the preceding, but in *all* previous revisions of the article (going backwards in time), reintroduced or reconstructed paragraphs and sentences are identified as such and the authorship information for the contained words is reconstituted.[9]

We implemented three additional variations of the described approach. For the first one, we use a Levenshtein distance metric to ignore word edits where less than three characters have been changed (mostly spelling mistake corrections). These words are counted as the same words and the authorship mark-up is not changed, so that small corrections don't lead to false authorship assignments. For the second variant, we ignore stopwords like "and" or "not" completely to not overestimate the contributions of editors who do a lot of grammar corrections but do not substantially contribute to the topic of the article. As a third alternative, we implemented a somewhat different approach, which uses the words to the left and the right of the word in question as its "coordinates" to identify changes. The text is divided into sequences of three- word-chunks which are compared to each other in every revision to find overlaps between them to identify newly added words. As no significant differences between the results of the described alternatives could be found, we will only discuss the baseline ArticleTree method's output in the following sections.

3.3 Evaluation of the algorithm

We tested both the part of the Wikitrust algorithm used to determine authorship and our proposed approach (baseline ArticleTree). We used the results of the openly available Wikitrust API[10] to compare accuracy of the found word-RevisionID-author tuple to our results. An accurately detected tuple means that the word

1. was indeed inserted into the article in the detected revision,
2. is the exact same word at that position (i.e. the word "and" in the introduction vs. "and" in a different paragraph),
3. was not reintroduced but inserted for the first time.[11]

As there is no reliable, evaluated gold standard for performing the task of word-to-author assignment in Wikipedia according to these exact conditions, we performed a manual evaluation. 250 word tokens were randomly selected from a sample of 45 randomly picked articles from the article namespace, which are not redirect or disambiguation pages. For each of these words, it was assessed if the Revision ID assigned by Wikitrust and WIKIGINI was correct, adhering to the above conditions. WIKIGINI assigned 59.2% of the words to the correct Revision ID, while Wikitrust did so for only 48.4% of the words. The difference is significant at p=0.001.[12] Splitting up the 40.8% (Wikitrust: 51.6%) errors made during detection, 36.8% (45.6%) detection results failed already at the first condition; the words were not added in that revision. For 2.4% (3.6%), the second condition was not met, and the remaining 1.6% (2.4%) that fulfilled the first two conditions were not introduced in that revision for the first time.

3.4 Discussion

Although the sample size is limited, the gain in accuracy of over 10% in the sample is significant and therefore indicates a notable accuracy improvement of our method not only in the sample, but in general, over the method employed by the current state of the art. As the share of around 40% false positives is not satisfactory, we are currently experimenting with refined methods using different kinds of tokenization, which seems to be one of the biggest sources for error.

4. THE GINI COEFFICIENT AS AN INEQUALITY MEASURE OF AUTHORSHIP

As pointed out for the Wikipraise tool, a simple listing of how many words the authors of a revision have written is not useful for the average editor to easily draw conclusions and does not allow for an aggregated metric that can be analyzed over time. We therefore calculate a "gini coefficient"[13] for the word authorship of every revision. A gini coefficient is a statistical measure for dispersion that is used to show inequality among values of a frequency distribution. One of its most common uses is to exemplify the inequality of per-capita- income among a nation's citizens. Its value range is in the interval [0,1], with value zero meaning complete equality (income for all citizens is the same), while value one means complete inequality (one citizen has all the income). It is mathematically based on the Lorenz Curve [5]. In our case, we use the gini coefficent to measure how equal the existing words in a specified article revision are distributed over the editors (in terms of authorship) that have been working on the article. We compute the the gini coefficient *G* as

$$G = \frac{2\sum_{i=1}^{n} iy_i}{n\sum_{i=1}^{n} y_i} - \frac{n+1}{n}$$

for values y_i, i=1 to n, which equal the amounts of words written by each editor in the analyzed revision, n being the total number of editors of the article.[14] The resulting coefficient gives an aggregated, simplified measure of how the authorship in the article is distributed in the current revision.

5. THE WIKIGINI TOOL

Using the generated data on word authorship and the calculated gini coefficients, we set up a tool to visualize their development over time for an article in an easily understandable manner.

5.1 Implementation

We implemented our WIKIGINI tool to be accessible online.[15] To depict the data for each article, we visualize the gini coefficient

[9]This is very relevant, as often, full or partial reverts to a previous state of the article are carried out, most frequently to restore the last unimpaired revision before a vandalism attack occurred, effectively re-introducing all of the words of that revision.

[10]Example call: `http://en.collaborativetrust.com/WikiTrust/RemoteAPI?method=wikimarkup&pageid=534366&revid=480773610`

[11]See Footnote 9. Wikitrust also addresses this issue and tries to find the original introduction. [1]

[12]Tested via a paired-sample T-Test. Results are available at `http://people.aifb.kit.edu/ffl/wikigini/`

[13]Also: "gini index" or "gini ratio".

[14]We assume a uniform probability distribution for y.

[15]`http://www.stud.uni-karlsruhe.de/~uebjn/index.php`

values over revisions in a graphical interface, as seen in Figure 2, using the JavaScript based solution by HighCharts.[16] It is possible to hover over the data points of a revision to see the gini coefficient and Revision ID. By marking a section of the graph it is magnified, allowing for closer inspection of the curve in that area. To decrease loading times and for comprehensibility, we show revisions in chunks of 1000, navigateable pagewise and starting from revision 1. Revisions are either placed along the x-axis at equal distances or proportional to the time that has passed between them (via the "xAxis" dropdown).

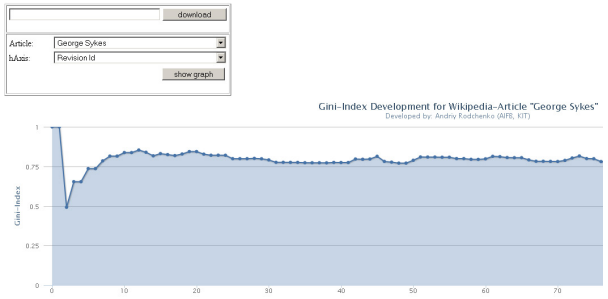

Figure 2: Interface of the Wikigini tool - Gini coefficient (original Article Tree) over time for article "George Sykes" (revision numbers on X-Axis, placed equidistant)

An article (currently from the english Wikipedia only) can be selected for analysis by entering the name into the appropriate text field and downloading it from Wikipedia with its complete revision history. It is then processed as described in Section 3.2, generating four coefficients per revision: the original ArticleTree method, ArticleTree neglecting stopwords and ArticleTree using Levenshtein distance, plus the 3-word-chunks method. The resulting curves are displayed in different colours and can each be toggled. Once an article has been processed, the results are stored. Already processed articles can be accessed through the article dropdown menu. As an additional feature, users can choose "Show last revision" to see the content of the last article revision as wiki syntax, where hovering over each word displays its detected author and revision ID of creation. This feature enables to easily identify the origin of a certain text passage.

5.2 Visualization examples and discussion

As mentioned in Section 3.2, the results of the four implemented methods didn't differ significantly. Accordingly, the gini curves for these methods show no mentionable differences either. In the following, we will thus show only the original ArticleTree curve to avoid visual clutter.

A common dynamic can be observed after the inception of an article: Following a short phase of volatility of about 50-100 edits, the gini coefficient usually becomes more stable for the rest of the article life, as can be seen in Figure 3, showcased by three articles. The figure also exemplifies how reaching this state of relative stableness can take more or less time, depending on how often the article is edited. This pattern intuitively makes sense, as the "founding phase" of an article naturally requires more rewriting and restructuring, until a satisfactory baseline article is set up, than after that period. Note, however, that the relative stableness after the founding phase can also differ greatly, as can be seen in Figure 3 as well: The coefficient of the article"Sergei Korolev" becomes very stable after about 60 edits, constantly decreasing afterwards, while for "Barack Obama" and "Lemur", it stays more volatile. Looking at these articles in detail reveals that the content of "Sergei Korolev" was a lot less contested from the beginning than the other two articles. The slow decrease of the coefficient also points to a rather "harmonic", slow build up of the article by many different editors in almost equal, rather small parts, a fact confirmed by inspecting the individual edits. Vandalism also plays almost no role in this early phase. This is indicated by the lack of spikes in the curve, which reliably signify vandals deleting all or almost all of the article, a very common vandalism form called "blanking".

For "Barack Obama" on the other hand, there are less individuals, each contributing larger chunks of content and there is more vandalism going on in this early stage already. Thus the coefficient fluctuates to a higher degree. Interesting is the edit of an anonymous editor at edit number 312, who rewrites large parts of the article, adding a lot of new content in the process, thus raising the coefficient by a good 0.08 points. He does so almost uncontested, as no revert takes places and the slow coefficient decrease afterwards indicates that his content is, if at all, very slowly rewritten or removed.

A similar, but much more distinct jump of 0.2 points occurs in the article "Antarctica" at edit number 757 (February 5, 2006) after its inception, as can be seen in Figure 4: In the course of 1.5 days and 20 edits, user "Mahanga" rewrites large parts of the article completely, deleting, shifting, but mainly adding content and constructing whole sections (interrupted only by two spelling corrections by another editor). The resulting coefficient of 0.906 slowly and monotonously decreases over the next 5000+ revisions as other authors make their contributions and changes, but don't essentially contest Mahanga's edits. This, as in the above example, indicates that Mahanga's entries are build upon by later edits, thus crucially impacting the article's development. Corresponding to this assumption is the fact that Mahanga is the top word contributor of the article at the time of his edits in question and stays in second position even until June 2012. Until this date, the sections History, Climate, Economy, Politics and Population introduced by him in the edits of 2006 as well as Biodiversity (a merger of his sections Flora and Fauna) survived, shaping the face of the article. For most of these, the content introduced in his 2006 edits still makes up about 30-60% of the sections. The section on Meteorites (a split-off from his Research section introduced in 2006) has been almost completely conserved over time.

Many more cases like the above can be found when analysing the coefficient development of articles, some mostly deleting, some adding and others replacing large parts of the content. The example shows clearly how a sudden and sustained change of the coefficient and its subsequent stabilization can indicate the act of a single editor leaving a strong mark on the article and fundamentally shaping its future development.[17] It is noteworthy that in our example, the changes performed by Mahanga were not discussed at all on the talk page before and after. Hence it is safe to assume that the idea and content came all from this one editor and were not a crowd decision. This is not to say there was no consensus on the content; it was just achieved ex-post, by not reverting Mahanga's edits instead of resulting from an ex-ante discussion and/or a collective building of the content through various editors. Such an increase (through one or a few editors) is not by itself a negative influence on article quality, but might well be in the case that only like-minded editors are active in the article at the time of the edits. Once the content has been in the article for some tens to hundreds of revisions, it can be very hard to change it (especially wording and writing style) by new editors, without very sound reasoning and references, although such requirements might not have been applied to the original content (like in the case of the Mahanga edits, were only one reference was added and no explanation for the edits was given). Hence, an abrupt and sustained de- or increase of the coefficient can act as useful signal in the article history to investigate the circumstances of how a specific section, paragraph or sentence came to be.

[16] http://www.highcharts.com

[17] In this case, one could as well have observed a steep rise in the number of characters. In many other instances however, editors replace a lot of content with their own, only marginally changing the length of the article. Vice versa, a jump in article length doesn't necessary point to a change in authorship concentration.

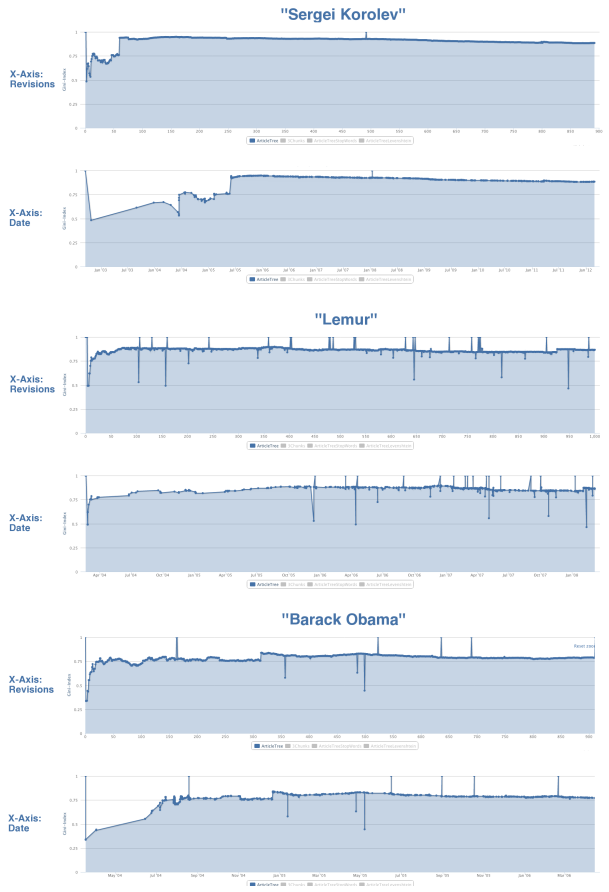

Figure 3: The inception phase of the articles "Lemur", "Barack Obama" and "Sergei Korolev". Upper rows with revision numbers, lower rows with date of revisions on the X-Axis.

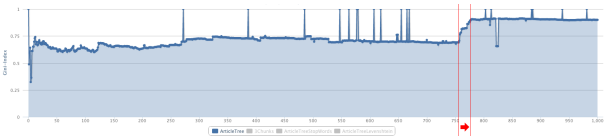

Figure 4: Inception phase of the "Antarctica" article, the coefficient increase caused by user "Mahanga" is marked up. Revision numbers on X-Axis.

5.3 Future work and planned tool extensions

The illustrations given in Section 5.2 are just some first examples of the analytical possibilities the gini coefficient provides. We plan to investigate further into correlations of gini coefficients with article templates like "featured", "POV" (i.e. biased article), "protected" and "maintained" , to name a few, as well as article categories, topicality, number of active editors in total, average edit frequency and various others. We plan to extend the WIKIGINI tool with information about more characteristics of the article to compare it with the gini coefficient development: character count, discussion page activity and percentage of word authorship by the top 10 editors, as well as mark up of reverts, vandalistic edits, and periods when important templates are present in the article. We also experiment with a feature to dive deeper into sections and paragraphs of an article to explore their gini coefficient development separately. Another experimental feature includes revision-to-revision comparisons of the distribution of words over editors to determine rank changes between the most contributing authors.

6. CONCLUSION

We presented a novel approach to detecting authorship of words in Wikipedia and showed that it significantly outperforms the baseline method employed by Wikitrust, reaching a 10% accuracy in our evaluation sample. This is achieved by addressing the minor degree of most edits, thereby reducing the necessary computation steps and uses of word-based text-to-text comparisons, which are the most fallible steps in the process, as our evaluation results indicate. Our motivation argues that the concentration of words to just a few authors can very likely be an indicator for a lack of quality and/or neutrality in an article, although the testing of this hypothesis remains to be delivered in future work. To provide an aggregated measure of the concentration, we calculate a gini coefficient for each revision of an article based on our word-author-assignments. The coefficient development over time in an article is visualized and provided online as an easily accessible and useful tool to investigate how the content of an article evolved. To provide some first application scenarios and analysis possibilities we presented examples where the changes in the gini coefficient give useful insights into differences between articles and may help to spot crucial events in the past evolution of an article. We are confident in the potential of the gini coefficient for analytical and statistical purposes and plan to expand the WIKIGINI tool to enable more elaborate visual comparisons of article dynamics.

7. REFERENCES

[1] T. Adler and L. Alfaro. A content-driven reputation system for the Wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, 2007.

[2] T. Adler, K. Chatterjee, L. Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *International Symposium on Wikis*, 2008.

[3] R. Burns and D. Long. A linear time, constant space differencing algorithm. In *Performance, Computing, and Communications Conference, 1997. IPCCC 1997., IEEE International*, pages 429–436. IEEE, 1997.

[4] F. Flöck, D. Vrandečić, and E. Simperl. Towards a diversity-minded Wikipedia. In *Proceedings of the ACM 3rd International Conference on Web Science 2011*, 06 2011.

[5] J. Gastwirth. The estimation of the lorenz curve and gini index. *The Review of Economics and Statistics*, 54(3):306–316, 1972.

[6] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 575–582, New York, NY, USA, 2004. ACM.