WikipediaViz: Conveying Article Quality for Casual Wikipedia Readers

Fanny Chevalier, Stéphane Huot, Jean-Daniel Fekete

WikipediaViz: Conveying Article Quality for Casual Wikipedia Readers

Fanny Chevalier *
Microsoft Resarch - INRIA Joint Centre

Stéphane Huot †
Université Paris-Sud & CNRS - INRIA.

Jean-Daniel Fekete ‡
INRIA

ABSTRACT

As Wikipedia has become one of the most used knowledge bases worldwide, the problem of the trustworthiness of the information it disseminates becomes central. With *WikipediaViz*, we introduce five visual indicators integrated to the Wikipedia layout that can keep casual Wikipedia readers aware of important meta-information about the articles they read.

The design of WikipediaViz was inspired by two participatory design sessions with expert Wikipedia writers and sociologists who explained the clues they used to quickly assess the trustworthiness of articles. According to these results, we propose five metrics for *Maturity* and *Quality* assessment of Wikipedia articles and their accompanying visualizations to provide the readers with important clues about the editing process at a glance.

We also report and discuss about the results of the user studies we conducted. Two preliminary pilot studies show that all our subjects trust Wikipedia articles almost blindly. With the third study, we show that WikipediaViz significantly reduces the time required to assess the quality of articles while maintaining a good accuracy.

Keywords: Wikipedia, Information Visualization, Encyclopedia, Collaborative Knowledge, Participatory Design.

1 INTRODUCTION

Wikipedia — the free online encyclopedia — has become one of the top ten most visited web sites in the world [13] with about 14 millions articles in 270 localized versions. This popularity mainly comes from its availability and coverage: Wikipedia is defined as "the free encyclopedia that anyone can edit", with "thousands of changes an hour". This fundamental Wikipedia concept has proved to be a good way to continuously increase the coverage, accuracy and up-to-datedness of information.

Conversely, this fast changing and volatile content is prone to unverified information. As a result, the question of quality and trustworthiness of the articles in Wikipedia has been heavily debated in the press [9]. As more and more people rely on Wikipedia, the cost of unreliable and incomplete information increases for society. Journalists that have no clues about the quality of articles content increasingly cite Wikipedia as a source on historical facts and figures and consequently can disseminate misleading information [14]. Helping Wikipedia readers, especially casual ones, spot questionable content that can provide erroneous or bad quality articles is thus becoming increasingly important.

To assess the maturity (and then the potential quality level) of an article, knowledgeable Wikipedia readers and contributors are used to finding clues in discussion pages, histories, and the visual appearance of articles. However, casual users are not aware of those clues, sometimes not even that quality may potentially be suspect [4]. One way to solve the problem would be to rank each entry. Recently, a non automatic system has been integrated in Wikipedia[1]. This system allows to rank articles according to their quality and importance relative to a project . But ranking articles is long and difficult and fewer than 1% of articles are currently ranked. Therefore, with the exception of a small number of ranked articles, there is no direct way of assessing the quality of articles on Wikipedia.

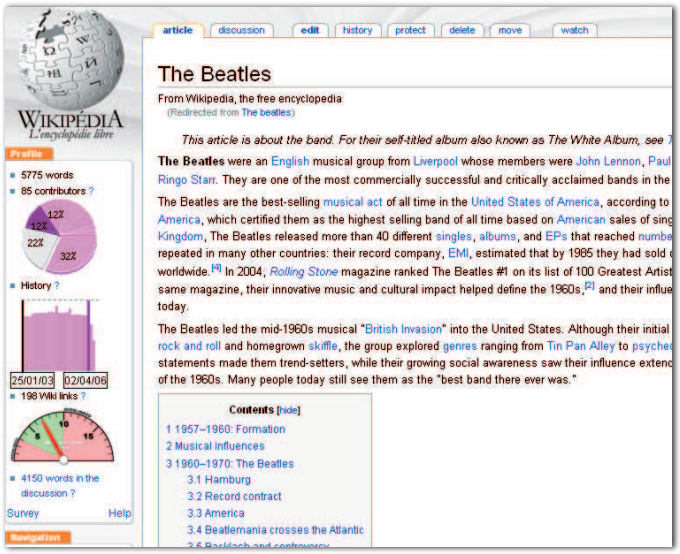

Figure 1: WikipediaViz visualizations revealing the history profile of a Wikipedia article.

In this article, we attempt to compensate for this lack of readily-available quality indicators by introducing five visual indicators to help Wikipedia readers judge the trustworthiness of the articles they read (Fig. 1). Because the quality of Wikipedia relies essentially on the editing and review process rather than the authority of the contributors, we designed these visualizations to improve the transparency of this existing process. Thus, as a first goal, our visual indicators should play an educative role as they reveal that Wikipedia articles are the outcome of a collaborative writing process. The challenge is then to raise the users' awareness of the editing process by displaying information about the history of the article in a simple and legible way. This would lead the readers to question the content and help them assess the quality of the Wikipedia articles. In particular, we are interested in *casual readers*: users who do not know the clues, are sometimes not even aware of quality issues, are not statistically-savvy and not visualization-savvy.

The first step of this work consisted for us to identify important clues about the quality of Wikipedia articles. We have worked with Wikipedia administrators, frequent contributors, and readers to establish objective metrics that they consider fundamental for drawing the profile of an article. We discovered that finding and interpreting data for these metrics is not straightforward and requires to be used to Wikipedia environment. To reveal and highlight these

*e-mail:fanny.chevalier@inria.fr
†e-mail:stephane.huot@lri.fr
‡e-mail:jean-daniel.fekete@inria.fr

[1] http://en.wikipedia.org/wiki/Wikipedia:1.0

criteria to the casual reader, we designed *WikipediaViz visualizations*, five visual indicators that represent these criteria in the left column of the standard Wikipedia layout (Fig. 1). These visualizations allow readers to grasp the profile of an article at a glance, instead of browsing and analyzing Wikipedia meta-data (discussion and history pages) on their own to gather significant clues about articles quality.

2 RELATED WORK

Trust is an important issue in online environments such as e-commerce, e-services or open collaborative web sites [8, 17]. In the special case of collaborative knowledge systems such as Wikipedia, their collaborative nature leads to strong arguments: critics say that anyone can write anything in an article, whether expert or not in the field; proponents answer that a non-expert can also improve or correct errors in an article that experts would not spend time correcting. This problem of quality and reliability on the articles of Wikipedia has become a topic of great interest.

2.1 Studying Wikipedia

Although many distrust Wikipedia, a qualitative study comparing Wikipedia and the traditional encyclopedia Britannica concluded that they are similar in terms of quality [9]. However, the choice of topics, the articles, and the method used to assess their quality (blind peer-reviewing) stirred up controversy. Beyond its results — which are quite difficult to prove formally — this study shows the difficulty of subjective assessment of the quality of Wikipedia articles.

There have also been quantitative studies that attempt to assess the quality of Wikipedia articles in a more objective way using metrics based on the meta-data associated with Wikipedia itself. Blumenstock simply uses the word count of an article as a measure of quality [2]. Lih [14] has suggested the number of edits and unique contributors of an article as measure for quality, where the number of unique authors reveals the "diversity" of an article and the number of edits its "rigor". Wilkinson et al. have demonstrated later that high-quality vs. non-featured articles have indeed substantially more contributors involved [25]. In [1], these two measures are combined to define the notion of "author reputation". Other single metrics based on the agregation of several indicators have been proposed to predict the quality of a contribution [7] or the trustworthiness of an article [6]. In addition the collaborative work in dedicated "discussion pages" — where changes are often discussed before being introduced in the article [21] — plays a critical role in the quality of articles [11, 23].

A common trend in these studies is that the metrics reveal social information that can be used as an indicator for assessing the quality of an article. It has been shown that revealing trust-relevant information to the users has an effect on the trustworthiness of the articles [12, 15], and three visual tools have been proposed aiming at enhancing the user and reader experience on Wikipedia.

2.2 Visualizing Wikipedia Meta-Data

Both the History Flow visualization [22] and WikiDashboard [15, 19] show the evolution of an article over time. The History Flow visualization relates the length of an article with the number of changes (characters added, removed or moved) and their authors. WikiDashboard provides an *article dashboard* that shows the weekly edit activity of the article, followed by a list of the corresponding main authors' activities; and the *user dashboard*, that displays the global weekly edit activity of the user, followed by a list of the pages the user has edited the most. However, WikiDashboard does not try to show cues about the quality of articles, so it is up to the reader to try to interpret the activity graphs as quality hints; this interpretation requires experience that casual users do not have.

While History Flow provides very accurate information about article histories and the users' contributions, it requires a large screen real-estate and some infovis education to be understood, as witnessed by our Wikipedia experts. In the same vein, Chromograms [23] is a visualization designed to understand the pattern of activity of prolific Wikipedia contributors. It is not aimed at casual readers but at investigating the sociology of Wikipedia.

All these visualizations allow for very precise analysis of articles and users, but they require a large portion of the screen real-estate to show the details. History Flow and Chromograms are not designed to read the article itself but to analyze it through visualizations of its meta-data: they are not targeted at casual readers.

2.3 Visualizations For Casual Users

Although visualization has mostly been used to allow a large quantity of information to be understood in a reasonable amount of time, it has also been used to show a small quantity of information very quickly. The "map of the market"[2] is one such visualization that allows understanding of the current state of the stock market in a matter of seconds, permitting finer investigations upon further interaction. This type of visualizations is also used on television, for sports such as baseball or tennis to help new viewers acquire the context. Well known small visualizations include Tufte's sparkelines [20] which he describes as "data intense, design-simple, word-sized graphics" and Hearst's Tile Bars [10].

Ambient visualizations [16] have been designed to be non intrusive and easy to understand so that they could be used by casual users. Notification systems [5] are an example of such ambient visualizations that are used to make the user constantly aware of events such as system status updates, email alerts, or chat messaging, but in a non-intrusive way. Security toolbars [26] are also designed to provide quick information such as warnings or certifications, using simple visualizations peripheral to the main window. In that sense, they all follow the concept of *Casual InfoVis* [16] that aims to be more "useful" and "utilitarian" for casual users, and to support different insights than traditional InfoVis systems. But so far, there have been very few attempts at proposing such simple and approachable visual representations that can help readers in better understanding Wikipedia and its possible issues. Adler and De Alfaro prototyped such a visualization based on their "author reputation" metric [1], that reveals the trustworthiness of portions of the text by highlighting them directly in the article. Even though this method is more approachable to casual readers than complex visualizations, highlighting parts of the text is intrusive and degrades the reading experience.

In summary, there have been efforts to define and assess the quality of Wikipedia articles but no solution has yet been proposed to expose it to the casual reader. Conversely, some tools allow visualizing and analyzing some aspects of Wikipedia data, but they focus on expert users or administrators; none of them is aimed at exposing quality issues in a simple way for the casual reader. There is a real need for tools that help readers to be aware of the volatile and changing nature of Wikipedia and to estimate the impact of this nature on the quality and trustworthiness of articles. As our targeted users are not expected to be experts in statistical analysis or information visualization, we propose to provide them with a quick

[2] http://www.smartmoney.com/marketmap

access to informations on the state and the evolution of an article by means of simple and relevant visual indicators. For this purpose, the first step of our work was to organize a participatory design session with knowledgeable Wikipedia users to identify how they deal with quality issues and what they would need to help them in this task.

3 Participatory design with Wikipedia contributors

We organized two participatory design sessions with knowledgeable contributors: four Wikipedia administrators, two heavy contributors and two sociologists studying Wikipedia (who are well aware of the various issues regarding its maintenance). We chose expert users because it would have been difficult for novice users to identify questions they don't even ask. From a user's point of view, the main goal of the first session was to identify the need of such "expert users" to deal with quality issues, especially when they browse Wikipedia as readers. The other goal was for us to identify the clues that they are using to quickly assess article quality.

The first discussions with participants strengthened the notion that the problem of quality assessment was an important topic in the Wikipedia community. Indeed, they all asked for tools to raise the awareness of quality issues for them as readers. These sessions resulted in two kinds of methods to solve this problem: objective and subjective measures. Objective methods exposed objective information on the main article page whereas subjective solutions tried to find aggregated scores to provide a quality rank of the page. Although both methods seem useful, we focused on objective methods because we felt they were more in the spirit of Wikipedia (neutral, verifiable and factual).

All the administrators and writers agreed that they no longer started to read an article without checking for some information: number of contributors, size and recency of the article discussion, number of recent edits and their sizes (e.g. typos fixed or paragraphs added). They all try to gather this information by looking at the pages related to the article (history and discussion) but not at the article itself. However, they also relied on the article style and format to note if it is "well-written", following the Wikipedia style and structure, having a reasonable number of references and internal links (hyperlinks to other Wikipedia articles) to verify the validity of the information and estimate its integration in the encyclopedia.

Most of the criteria exposed by expert subjects were validated by the studies on Wikipedia we mentioned before. Moreover, they were also closely related to the maturity of articles, suggesting that the maturity of an article plays an important role for them in assessing its quality. The correlation between maturity and quality in Wikipedia has been demonstrated in [3] by analyzing several classes of articles (stub, normal, good and featured) according to simple metrics (number of words, headlines, images and links). To estimate and highlight the maturity of articles, the authors also discussed the idea of providing an aggregated maturity level using their metrics. However, their computed value exhibits a high standard deviation across articles and does not seem reliable alone without human supervision.

Five Metrics for Maturity/Quality Assessment

Using the results of these sessions and the related work about quality in Wikipedia, we have defined five objective metrics that are currently used by expert users to decide their initial opinion of an article.

Word Count. The article length is a simple indicator that gives insight as to the amount of information contained in the page. Intuitively, we can assume that featured articles are long (detailed and complete). But it is less obvious that long articles are good. Blumerstock has shown that the word count is one of the most effective metrics to differentiate featured articles from other articles in Wikipedia [2], but he noticed several counterexamples that prevent using it alone as a quality measure. We could have used other measures computed through standard natural language processing techniques, e.g. to identify misspelled words or assess the diversity of the vocabulary. However, these measures are hard to interpret for casual readers, expensive to compute and were not asked during our participatory design sessions with Wikipedia experts so we did not include them.

Number of Contributors and **Rate of Contribution**. The number of distinct contributors and the length of their contribution that remains in the current version of the article provide several clues for experienced readers. It can show whether there is one major contributor or several, and the approximate distribution of the contributions. It is a good indicator of the involvement of the community in the article and its topic. Moreover, it has been shown that high-quality articles in Wikipedia are distinguished from the rest by having a larger number of distinct contributors [24, 25].

Number and Lengths of Edits. The number of edits of the article and the lengths of the versions are also important clues for expert readers. Ideally, a page reaches its highest quality level when it has been edited several times to be completed and corrected [25]. Furthermore, the length of the contributions shows if they are major edits (new or updated content) or simply typo corrections. Finally, this metric provides more clues when taking into account temporal aspects (as described in [19, 22]): few edits over a long time usually denote that the article has reached a good level of completeness (relative to its length), or that it is not a hot topic among the Wikipedia community. Conversely, several edits in a short time can denote conflicts or vandalism (edit wars) that sometimes occurs in Wikipedia. It can also reveal a resurgence of interest in the article topic, such as in connection with current events (*e.g.* the election of a new president triggers updates to his/her page.)

Number of References and Internal Links. Revealing the source of information has been recognized as an important factor influencing trustworthiness [18]. The Wikipedia standard for writing articles encourages using links to reference sources to quickly verify or refine information. Consequently, knowledgeable readers are suspicious when important facts are not supported by references in an article. In the same manner, they also observe the number of links to other Wikipedia articles (named *internal links* or *wikilinks*) that an article contains. This informs them of the integration of the article in a pool of articles or in a category. The Wikipedia Manual of Style[3] also encourages the use of links to other articles to help users in following "their curiosity or research" on a topic. Of course, there is not an ideal number of internal links for a given article. However, a small number can suggest a relatively recent article that is not yet well integrated into the encyclopedia, most of the time a "copy and paste" from some other source or one that has not yet been reviewed by knowledgeable contributors. The article is then qualified as *underlinked*. Conversely, an *overlinked* article can drive the reader to irrelevant articles. In the Wikipedia guidelines, the first mentioned criterion is that an article is overlinked when more than 10% of the words are contained in links, with some notable exceptions.

Length and Activity of the Discussion. The discussion page re-

[3] http://en.wikipedia.org/wiki/Wikipedia:ManualofStyle

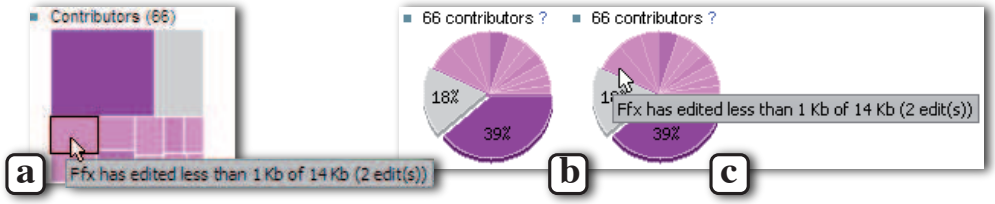

Figure 2: Interactive visualization of authors' contribution sizes: treemap (a) and pie chart (b) & (c).

lated to an article is used by contributors and administrators to exchange ideas on the content and writing style of the article or for planning its evolution. They allow contributors to communicate with each others to come to an agreement about controversies, contentious parts of the article, or simply its organization. Cooperation and coordination have a strong impact on the quality of an article [21, 24, 25] and frequent readers always have a look at the discussion page when it exists. Excluding the content of the discussion, the two major clues they observe are its length and its recent activity. The length of the discussion informs them about potential controversies or consensus that could have been raised and discussed along the lifespan of the page. Its recent activity provides the same clues on the current (or recent) version of the article.

4 WikipediaViz Visualizations: Design Iterations

Although the metrics we proposed give clues about the quality of an article, they can not be used as they are by readers for two reasons: some of them could only be gathered through extensive and tedious navigation using the standard Wikipedia interface; this is out of reach of casual Wikipedia users. Other metrics required heavy computations on all the Wikipedia revisions. Therefore, we have modified the Wikipedia interface to add five visualizations based on theses metrics for depicting the profile of an article (Fig. 1), inserted as Dynamic HTML objects or images in the left panel, under the Wikipedia logo.

We designed the visualizations to be small, simple and expressive. In terms of interaction, we decided to mostly provide tooltips for details but not to augment the visualizations with navigation capabilities that would distract the users from reading the article and that would require some documentation.

Finally, we went through two iterations to achieve a usable interface since the first version revealed unexpected behavior from casual Wikipedia users.

4.1 Word Count

We designed no specific visualization for the word count metric as it only consists in a single value. We merely show it textually (Fig. 3).

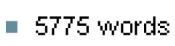

Figure 3: Word count of an article.

4.2 Authors Contributions

The visualizations we propose aim at associating the number of authors, their contribution rates and the number of edits they made in the same view.

First iteration. For each article, the contribution of all its authors was visualized as a Squarified Treemap [3] (Fig. 2a). The whole square represents the current length of the text and each nested square represents the contribution of a unique author. The area of a square is proportional to the amount of text that remains in the displayed version of the article. Squares are colored according to the number of contributions made by the corresponding author to the article with lighter color for few contributions and darker for more. For example, a small and dark square indicates a frequent contributor relative to the article with a small number of characters left in current version of the article. To avoid clutter in the visualization, small contributions that would take fewer than a few pixels (above 1/100th of the total area) are aggregated and displayed as a single gray square.

Our initial study revealed that casual users had difficulties understanding the Treemap so we changed it in the second iteration.

Second iteration. We redesigned the authors' contribution visualization to be as simple and evocative as possible, following the design principle of *Casual InfoVis* indentified in [16]. We replaced the treemap by a pie-chart (Fig. 2b), which is a well-known graphical representation that is easier to understand and interpret. The visualization displayed the same information as the previous one, with the same approach of space distribution and color code.

We provided interactions in the visualization to support some exploration. More details about the contributions are obtained by moving the pointer over a slice in the pie chart (Fig. 2c): a tooltip is displayed that shows the author's name, the aggregated size of the contribution and the number of edits.

With the standard Wikipedia layout and tools, one needs to browse the entire history of a page to count authors and edits; computing the contribution rates manually is practically impossible. It requires one to compute the author of each character for each article. The visualizations we propose associate the number of authors, their contribution rates and the number of edits they made in the same view. In particular, they quickly show when an article has been mostly written by one author or several. Though quality articles tend to be written by more than one author (large contributions come from a few authors) and edited by many to fix typos and improve stylistic or structural issues (several minor edits). Conversely, the pie chart also reveals articles that are written by a small number of contributors. Although this is not a proof of low quality (the article could have been written by a few experts of the topic), this usually indicates stubs or pages that have not been extensively read and reviewed and that are prone to contain some mistakes.

4.3 Article Timeline

Figure 4 shows the timeline that visualizes the evolution of the article length. The time range is displayed under the graph with the rightmost bar corresponding to the current day; the length is displayed as bar height. The graph resolution (time scale) is computed so that it fits the visualization area with a lower limit of 3 pixels for

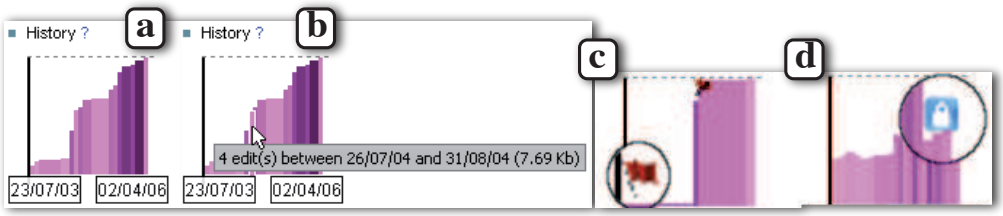

Figure 4: Plot graph of contributions lengths across time.

a time period. When several changes occur during a shorter period, the average is displayed. Conversely, when there are no changes during long periods, the last length is continuously displayed until a change occurs. The bars are colored according to the number of edits that have occurred during the time period. To provide more details about the contributions, when the user moves the pointer over a time period, a tooltip displays the starting and ending dates of that period and the number of edits that occurred (Fig.4b).

The history pages of an article allow one to retrieve older versions of an article and to see an overview of its evolution. However, it is hard to see stress — such as frequent edits in a short amount of time — and accidents — such as the split of an article in two or restructuration (see Fig. 6d) — that our visualization quickly shows. With this graph, maturity can be seen as a stabilization of the slope of the length at the end of the article and evolution stages are visible as sharp changes in the timeline. The color (expressing the number of edits) allows the visualization of high activity. So-called *edit wars* — when one user changes a portion of an article and another rewrites the same portion immediately over and over — can also be quickly inferred by the most knowledgable users as they most often correspond to periods with a stable size and a dark color.

The article timeline graphic was improved in the second iteration to present additional information by means of small icons (Fig. 4c and d). These markers show important events that have occurred on the article lifespan: banners and protections. In Wikipedia, banners can be placed on an article to call the attention of the reader about several properties of the article. Common banners notify that the article is a stub or that it does not provide enough external references. In our visualization, banners that have been placed in an article are reported with a red flag on top of the corresponding bar in the timeline. A crossed flag indicate that a banner has been removed (see Fig. 4c). A more critical event is protection. Articles that are prone to controversies or vandalism can be semi-protected (only registered users can edit) or protected (authorization should be granted by an administrator to edit). A protection event is reported on the timeline by a lock icon (Fig. 4d) and a crossed lock is displayed when the article is unprotected. These icons allow identifying important events that are difficult to find otherwise.

4.4 Internal Links Meter

To display the internal links number, we designed a "gauge" visualization that shows the density of internal links per page (percentage of words that are contained in internal links) and reveals *underlinked* or *overlinked* articles (Fig. 5). The number of internal links contained in the page is also displayed as text. The colors in the background show three zones corresponding to underlinked, fair and overlinked densities.

The rightmost red zone (higher than 10%) indicates an overlinked article. The 10% value comes from the Wikipedia style guidelines that consider an article as overlinked when it exceeds that value.

Figure 5: Internal links meter (Underlinked, Fair, Overlinked).

The leftmost red zone indicates underlinked article (fewer than 3%). We determined this 3% threshold by analyzing the featured articles of Wikipedia as they are supposed to comply with the Wikipedia guidelines. Our analysis of the French Wikipedia (12 June 2007) showed that about 80% of the 345 featured articles have an internal link density of 3-10% and that only 4% of them have a density higher than 10%. Consequently, we deduced that a threshold of 3% was representative of underlinked articles: an article can be considered as "fairly" linked when it has an internal link density in the 3-10% range (the green zone). The gauge is large because it both shows the actual value and the range of good values according to Wikipedia; a bare number would be impossible to interpret by casual readers without this context information.

The standard presentation of Wikipedia articles provides a rough overview of the density of internal links (links are highlighted in blue) but does not provide a quick way to identify underlinked and overlinked articles. Furthermore, when an article is longer than the browser window, it requires scrolling to see all the links whereas our indicator displays the information in a concise way on the top.

4.5 Discussion Length and Activity Indicator

Readers can easily access the discussion page of an article in Wikipedia. However, beginners and casual readers rarely look at the discussion page [4]. To highlight the existence of recent discussions, we inserted a simple indicator at the bottom of our profile visualizations. If there is a discussion page related to the current article, its length is displayed in textual form. Furthermore, if new entries were added to the discussion during the last two weeks, the indicator becomes a link to the discussion pages. This highlights recent activity of the discussion and encourages the reader to visit the discussion.

4.6 Typical Article Profiles

The documentation regarding all the visualizations can be accessed by clicking on the small question marks next to them. This documentation shows several examples of typical article profiles to better explain the interpretation of WikipediaViz. In this section, we show an excerpt of them extracted from of our working-copy of the French Wikipedia database (July 2008).

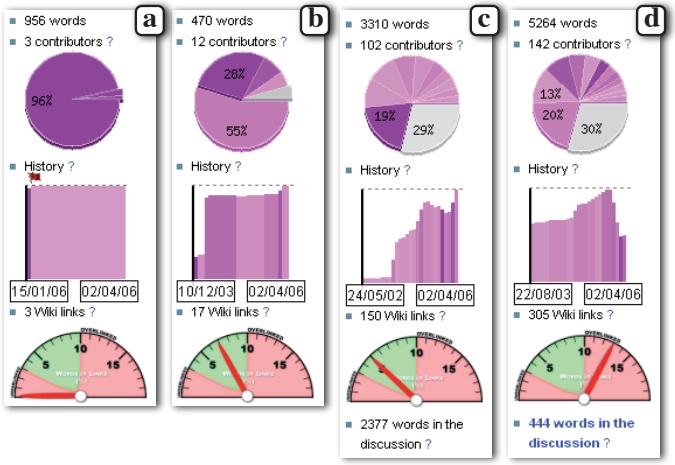

Figure 6: Examples of article profiles: *stub* (a), *good start* (b) and *mature* (c) & (d).

When looking at the 4 article profiles on Fig. 6, we quickly notice that the first is more recent than the others and that it has a small number of contributions. It shows the typical profile of a *stub*: created a few months ago, with few contributors among which one has done most of the content, a flat timeline, few internal links and no discussion. In most cases, these articles are also very short; this one contains around 1000 words. Without the visualizations, a casual reader could assess that this article length denotes a certain level of maturity whereas the visualizations show the contrary.

The b, c and d profiles show more mature articles. The b article is reaching maturity, with one predominant contributor, a length that has reached a plateau and a good link density. One noticeable detail is that it has no discussion page. We can consider that this article is a good start. Finally, c and d profiles are from mature articles. They both have a large number of contributors, but they differ in the shape of their timeline, their internal link density and their discussion activity. The c profile has reached a length plateau, has a good link density. There are some entries in its discussion page but no recent activity. It denotes that the article has reached stability. Conversely, the d profile shows a more stressed length timeline, a high link density and recent activity in the discussion. Even if this article can be considered as mature, these indicators reveal that the article is still changing and the reader should take this information in consideration. When reading the discussion page, we noticed that contributors are working on a reorganization to make it shorter and more focused.

5 User Studies and Discussions

We conducted three studies to evaluate the effectiveness of the WikipediaViz visualizations to help users gain a first estimation of an article quality.

5.1 Pilot Studies

We focused these preliminary studies on the influence of our visualizations on the confidence the users have in articles. However, these two studies were conducted before quality and significance assessment of articles was integrated in Wikipedia. Consequently, we missed a classification of articles as a reference.
We asked subjects to answer several factual questions about a sample of articles, with or without WikipediaViz. Answers were limited to 3 choices and the subjects had to give a level of confidence for each answer (from "Not sure at all" to "Very confident").

Post-experiment interviews showed that most of the subjects did not question the quality of articles. Furthermore, they admitted that even when they had noticed the visualizations, they did not try to use them and focused only on the question to answer. Consequently, the visualizations would not have helped them in estimating their level of confidence in the article.

As the first study showed us that the visualizations alone could not warn the readers of potential quality issues, we conducted a second study where the subjects were informed of the problem to encourage them to use the visualizations. We found no significant effects of the visualizations on the user performance but could not draw any conclusion due to the questionable choice of the articles.

5.2 Controlled Study

In the third study, we used the classification recently introduced in Wikipedia[4] as a reference to compare the quality rank of articles to the quality we asked the users to assess.

Experimental setup. 24 unpaid subjects, 19 males and 5 females, aged from 22 to 62 served in a within-subject experiment. It was conducted at a science library and at a science museum. Before the experiment, we asked subjects to answer a short questionnaire about their background in computer literacy, internet and electronic encyclopedias. 14 of the subjects were casual Wikipedia readers, 2 did not know of Wikipedia and 8 were frequent readers. No contributor took part in the experiment. We gave each subject a 1 page manual containing an introduction to Wikipedia — making them aware of the potential quality issues —, a detailed explanation of the visualizations and some typical WikipediaViz profiles to help them interpreting them. This training phase took 5 to 10 minutes.

We asked the subjects to assess the quality of 24 articles, 6 for each level of quality among Featured, Good article, Good start, and Stub (Quality factor). The articles were presented with different interfaces (Technique factor): Wikipedia interface (A), and WikipediaViz visualizations only (V), i.e. the article content was hidden; only its title was displayed. We revealed the content of the article after the answer. The experiment consisted in two blocks: the baseline condition block, in which articles were displayed as in Wikipedia (A), and the WikipediaViz block, in which articles were displayed without their content (V). Quality and Technique were counterbalanced across subjects using a Latin square. We logged the quality estimation of the user and the Time to perform the estimation.

Articles were offline versions of the July 2008 French Wikipedia. External hyperlinks and the Wikipedia search engine were disabled to prevent browsing other pages. Each trial was limited to 1 minute to limit the total experiment time but also to force subjects to focus on their first impression.

Hypothesis. Our hypothesis was that the visualizations could help the reader assess the quality of an article, both in time and in precision. It would be visible by a correlation between the technique and both the time and precision. We define Precision (P) as the distance between the quality of the article (from 1 for Featured to 4 for Stub) and the users' answer.

Results and discussion. Analysis of variance shows no significant effect of Quality and Technique on P. The graph of P by Quality and by Technique (Fig. 7a) shows that P is very similar and

[4] http://en.wikipedia.org/wiki/Wikipedia:1.0

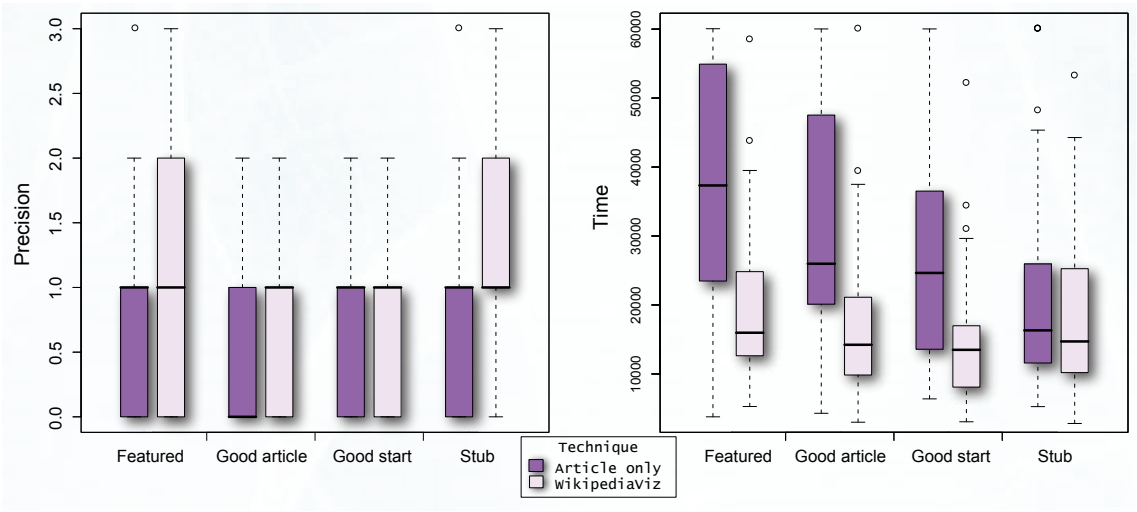

Figure 7: (a) Precision of the answer and (b) Performance, Time by Technique and Quality.

small under the different Techniques and Quality levels: the quality assessment is alway good with or without the visualizations. Even if it does not validate our hypothesis, this result shows that the WikipediaViz visualizations alone are expressive enough to assess the quality of articles at a glance with a similar precision as when seeing the full article at length.

Analysis of variance reveals a significant effect of Technique and Quality on Time ($F_{1,23} = 31.60$, $p < 0.0001$ and $F_{3,69} = 15.52$, $p < 0.0001$). Post-hoc means pairwise comparisons (Student's t-test, $\alpha = 0.050$) indicated that the time required to achieve the task was lower when the visualizations only were visible (V, mean=16.8s) than when the content only was visible (A, mean=29.7s). Post-hoc analysis for Quality (Tukey HSD, $\alpha = 0.050$) showed that Featured and Good articles require more time to be estimated than Good start and Stubs (Fig. 7b). We interpret this result by considering that articles of lower quality (short contents and specific visualizations) are easier to recognize than high-quality ones that require a deeper analysis.

The results of this study do not fit completely our hypotheses. However, it demonstrates that WikipediaViz reduces the time required to assess the level of quality by a factor of about 2 and without loss of precision (the precision of the estimation has a median value of 1 in each visualization condition).

Subjective evaluation. We asked the subjects to answer a short questionnaire after the experiment to gather feedback on the visualizations. All of them found that displaying summarized information about the editing process of an article was useful. They all answer that the rank they gave using the visualizations only was almost always the rank they would have given once the article content was revealed. They agreed that the number of contributors and their contribution rates are very useful to assess the quality of the article as it depicts the involvement of the community. They all agreed that a cooperative process leads to an improvement of the quality and relied mostly on this contribution rates to make their opinion. Surprisingly, most of them admitted that they did not pay attention to the length and recent activity of the discussion and whether it can be an important indicator revealing that a real process of collaboration and coordination of the contributors has occurred. They argued that one can write huge comments saying nothing interesting and did not actually trust this metric.

Also, several participants mentioned that it would have been interesting to see the how many times an article has been visited since it shows the popularity of the article. We agree that it is a useful information since mistakes have fewer chances to remain in an article that is often visited than in one that is seldom viewed. Other metrics have been mentioned by our participants: number of bibliographic citations, number of figures, etc. Some participants wanted a simple aggregated measure such as "good" or "bad" they could blindly trust but had no idea on how it could be obtained.

Finally, several subjects mentioned that they learned much more from the visualizations than by looking at the articles, but it took them time to understand what the visualizations were showing. This means that the use of WikipediaViz needs some practice, but that casual users are able to learn how to use them effectively.

6 Implementation Issues

WikipediaViz is implemented in PHP and integrated as a plugin in the Mediawiki system distributed and maintained by developers for the Wikipedia Foundation. It relies on the standard Wikipedia database with additional tables computed to quickly visualize the timeline and author contributions. These tables are currently stored in our static copy of the Wikipedia database and required three weeks of computation using the history data made available by the Wikipedia Foundation. If our visualizations are to be included in a production version of Wikipedia, the incremental computation of this information would take a negligible time and relatively small space. However, considering the current load of Wikipedia servers, this may be an issue to consider.

The generation of visualizations in web pages mainly uses HTML boxes, except for the pie-chart that is an image generated on-the-fly. These images are cached using the same mechanism as the one used to cache the web page generated from the Wikipedia database.

The code is available as free software and the experimental version is also available for feedback from the Wikipedia community, initially for a limited number of Wikipedia administrators and, after a test period, to the rest of the Internet.

7 Conclusion

In this article, we have described the iterative design of WikipediaViz: five casual visualizations aimed at improving the standard interface of Wikipedia for casual readers. These visualizations have been designed after two participatory design sessions

conducted with Wikipedia administrators, prolific contributors and sociologists. They show measures that are important for assessing the quality of articles but difficult to gather using the standard interface and that casual readers do not even know. We show that they significantly improve the time required for assessing the quality of articles with no effect on the precision of the assessment which is good.

But our experiment results are not definitive and will require more studies but they already reveal important issues about blind trust in Wikipedia that we had not anticipated. In that sense, we believe WikipediaViz can contribute to a better use of Wikipedia and address some of the current issues such as the proliferation of banners and the painful navigation required to understand the profile or articles. However, novice users may still not be able to interpret the visualizations or even see their significance because most of them are not aware of the issues raised by the Wikipedia editing process.

Our future work on WikipediaViz will focus on two points. First, now that WikipediaViz has proved to be useful, we have started to implement a live version where the metrics values are computed on the fly. The infrastructure needed to maintain an up-to-date version of Wikipedia is large and complex so we did not want to do so before we were sure it was effective. We hope that it will be a first step to their inclusion in the standard Wikipedia system.

Finally, despite the proved effectiveness of our visualizations, more longitudinal studies are required to understand their acceptability and their usage by casual users. More generally, it connects with a deeper reflexion we have on the fact that simple casual visualizations can be effectively used to enhance the user's experience while accessing dynamic contents, but the evaluation of the benefit they provide remains a challenging task.

ACKNOWLEDGEMENTS

Thanks to the Palais de la Découverte and the Cité des Sciences et de l'industrie. Thanks to Julien Levrel for his help and to the Wikipedia community for its support. This work was partly supported by the French ANR Autograph project.

REFERENCES

[1] T. Adler and L. de 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, New York, NY, USA, 2007. ACM Press.

[2] J. Blumenstock. Size matters: word count as a measure of quality on wikipedia. In *WWW'08: Proceedings of the 17th international conference on World Wide Web*, pages 1095–1096, New York, NY, USA, 2008. ACM.

[3] M. Bruls, K. Huizing, and J. J. V. Wijk. Squarified treemaps. In *In Proceedings of the Joint Eurographics and IEEE TCVG Symposium on Visualization*, pages 33–42. Press, 2000.

[4] S. L. Bryant, A. Forte, and A. Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *GROUP'05: Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, pages 1–10, New York, NY, USA, 2005. ACM Press.

[5] J. M. Carroll, D. C. Neale, P. L. Isenhour, M. B. Rosson, and D. S. Mccrickard. Notification and awareness: Synchronizing task-oriented collaborative activity. *International Journal of Human-Computer Studies*, 58:605–632, 2003.

[6] P. Dondio, S. Barrett, S. Weber, and J. Seigneur. Extracting trust from domain analysis: A case study on the wikipedia project. *Autonomic and Trusted Computing*, pages 362–373, 2006.

[7] G. Druck, G. Miklau, and A. McCallum. Learning to predict the quality of contributions to wikipedia. In *WIKIAI 08*, pages 7–12, 2008.

[8] B. J. Fogg and H. Tseng. The elements of computer credibility. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 80–87, New York, NY, USA, 1999. ACM Press.

[9] J. Giles. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901, December 2005.

[10] M. A. Hearst. Tilebars: visualization of term distribution information in full text information access. In *CHI'95: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 59–66, New York, NY, USA, 1995. ACM Press/Addison-Wesley Publishing Co.

[11] A. Kittur and R. E. Kraut. Harnessing the wisdom of crowds in wikipedia: quality through coordination. In *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 37–46, New York, NY, USA, 2008. ACM.

[12] A. Kittur, B. Suh, and E. H. Chi. Can you ever trust a wiki?: impacting perceived trustworthiness in wikipedia. In *CSCW '08: Proceedings of the ACM 2008 conference on Computer supported cooperative work*, pages 477–480, New York, NY, USA, 2008. ACM.

[13] R. Lee and T. Bill. Wikipedia users. Pew Internet and American Life Project, 2007.

[14] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *5th International Symposium on Online Journalism*, 2004.

[15] P. Pirolli, E. Wollny, and B. Suh. So you know you're getting the best possible information: a tool that increases wikipedia credibility. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 1505–1508, 2009.

[16] Z. Pousman, J. Stasko, and M. Mateas. Casual information visualization: Depictions of data in everyday life. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1145–1152, 2007.

[17] B. Shneiderman. Designing trust into online experiences. *Commun. ACM*, 43(12):57–59, December 2000.

[18] B. Stvilia, M. B. Twidale, L. C. Smith, and L. Gasser. Assessing information quality of a community-based encyclopedia. In *Proceedings of the International Conference on Information Quality - ICIQ 2005*, pages 442–454, 2005.

[19] B. Suh, E. H. Chi, A. Kittur, and B. A. Pendleton. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. In *CHI'08: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, pages 1037–1040, New York, NY, USA, 2008. ACM.

[20] E. R. Tufte. *Beautiful Evidence*. Graphics Press, 2006.

[21] F. B. Viegas, M. Wattenberg, J. Kriss, and F. van Ham. Talk before you type: Coordination in wikipedia. *Hawaii International Conference on System Sciences*, 0:78a, 2007.

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

[23] M. Wattenberg, F. Viégas, and K. Hollenbach. Visualizing activity on wikipedia with chromograms. In *INTERACT 2007*, pages 272–287, 2007.

[24] D. M. Wilkinson and B. A. Huberman. Assessing the value of cooperation in wikipedia. *Firstmonday*, 2007.

[25] D. M. Wilkinson and B. A. Huberman. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, New York, NY, USA, 2007.

[26] M. Wu, R. C. Miller, and S. L. Garfinkel. Do security toolbars actually prevent phishing attacks? In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 601–610, New York, NY, USA, 2006. ACM Press.