On the analysis of Wikipedia activity through time

Nuno Silva, Daniel Gonçalves
INESC-ID / Instituto Superior Técnico / University of Lisbon
Lisbon, Portugal
e-mail: nunojsilva@tecnico.ulisboa.pt, daniel.goncalves@inesc-id.pt

***Abstract*—Wikipedia articles see bursts of update activity whenever a topic is of more interest to the community or has somehow become controversial. Analyzing when and what changes are made can, thus, give us an idea of how the community feels about particular subjects. In this paper we present PopCulture, a system that provides a visualization of Wikipedia's edits that allows us to reflect on how different subjects are perceived by people over time and, by comparing articles from different language wikipedias, find regional and cultural differences of interest and perception. A set of user studies shows that, indeed, users are able to use PopCulture effectively and efficiently to find such trends and differences.**

Keywords-Visualization,Wikipedia, Cultural Differences

I. INTRODUCTION

Wikipedia allows users to submit changes to the content. Since previous versions are stored on the site, accessing them makes it possible to collect information on users, topics and their importance and relevance through time. Behaviors and trends assessed from the Wikipedia edition history can sometimes become a proxy for public opinion. Indeed, since Wikipedia articles are a collaborative effort, they reflect the opinion of a community of users and not of a single person. Not all users may be in agreement, of course, but that is often made apparent by the number of edits and reverts a page suffers. A polemic issue will generate much more activity than a consensual one. Subjects people have stronger feelings about or on which their position is more extreme can show signs of vandalism. Levels of activity or significant edits can also point to when some new information about a subject became apparent to users, who then promptly update the articles. It can be argued to what extent are Wikipedia contributors representative of the population at large. Nevertheless, in the general case, they are immersed in their respective societies and such editing patterns will be linked, in some way, to what is occurring in them.

Alas, Wikipedia shows only the most current revision of an article, hiding all the aforementioned information. We use infovis techniques to highlight relevant patterns in the data. It also allows users to compare two articles, possibly from different Wikipedias, nourishing comparisons between the different communities of Wikipedia editors and of speakers of the involved languages. This may highlight different perceptions on a same subject, or different times when new information became apparent in different contexts.

Visualization of Wikipedia data has already been the focus previous research. Brandes et al. [2] propose network visualizations, focused on authors, not topics, but constrained with respect to the conveyable information, which, except for Brandes et al., does not include evolution through time. WikiDashboard [5] and WikipediaViz [3] provide several metrics, some of them along with their variation along time, integrated in the display of the content of a single Wikipedia article version. Omnipedia [1], unlike the other works, explicitly focuses on the comparison of the coverage of topics, and of relationships between topics, across the different editions of Wikipedia operated by the Wikimedia Foundation. History flow [6] present content-focused visualizations, where their evolution through time is a key part of the screen, being the only of the surveyed works that effectively enable users to analyze changes in the depicted information through time using the main element of the visualization.

None of these works has focused on the comparison of the evolution through time of different versions of the same article, across Wikipedias.

II. THE POPCULTURE VISUALIZATION

Our system works for any of the different Wikipedia versions. The site language is used to choose the list of offensive, biased and good words used in the computed metrics (currently for English, French, Portuguese and Spanish). We based our metrics on Mola-Velasco's [4] work: article size, authorship, Quality (higher for more readable, information-rich articles), Vandalism (higher for articles with less readable/informative or offensive words) and Controversy (rich in good words while using biased words).

Central to the visualization (Fig. 1) are two plots along horizontal timelines, which encode information about edits, marked using small rectangles (whose height is proportional to the edit impact) and color coded according to their kind (blue for reverts, red for deletions and white for regular edits), and edits to the corresponding talk page, marked with green circles. The user can pan and zoom on the timeline, and the text of a particular revision (with highlighted differences between consecutive revisions) can be shown on a sliding side pane. Hovering over the plot displays a tooltip with information about the metrics at that point. Both plot height and color can be configured to depict the different metrics. Color can also be used to represent the authorship of revisions. The upper plot and a lower plot allow users to compare either different metrics for the same article or to compare two different articles, possibly from different sites, on the same screen. Each of these parts can be set to convey information for the entire article (using only one plot) or for each of its sections (rendering one plot by section).

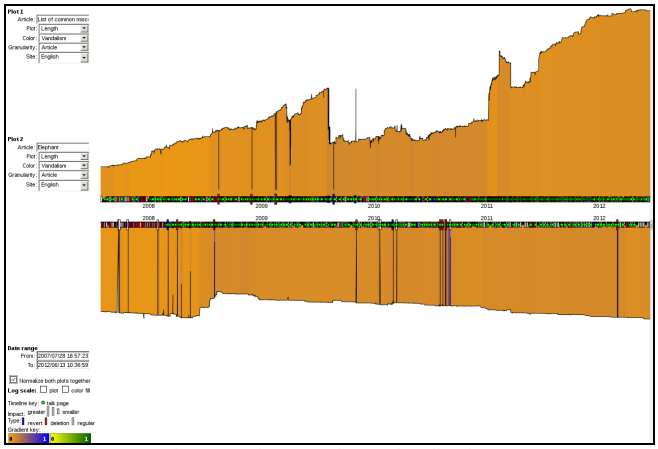

Figure 1. The Popculture Visualization

III. Case Studies

We used the visualization to explore some articles, in an attempt to discover patterns, similarities and differences. In the English and Hungarian articles on former Hungarian president Schmitt we could, by focusing on 2012, find two activity bursts, on January and April. This not only shows there is similar activity in both articles, but is also an example of a correlation between Wikipedia article activity and events outside Wikipedia: on 11 January 2012, a magazine accused Schmitt of plagiarizing his PhD dissertation, explaining the first burst; After the scandal played out, Schmitt announced his resignation from the office of President of Hungary on 2 April, matching the second burst. That the English language article results in a pattern similar to the Magyar version isn't surprising, as the English version of Wikipedia is it's a more international, encompassing, version. People from all nationalities edit it and it is a sort of superset of all editions. On the other hand, other language editions are edited only by speakers of those languages who are, thus, much more uniform both in geographic location and culture. We can see that by comparing the English and Portuguese articles: the claims of plagiarism did not translate into heavy activity bursts in the Portuguese article, although some editors promptly updated it. This can indicate that the Pál Schmitt scandal was given much less importance in Portuguese-speaking countries.

This type of bursts, triggered by events outside Wikipedia, can also be observed on the Portuguese article on "Fernando Nobre", displaying activity spikes that did not occur in the English version (hinting at their mostly regional relevance). Fernando Nobre was involved in a minor scandal around his possible running for elected president of the Assembly of the Republic. Visualizing the entire lifespan of the articles on "Fernando Nobre" from the Portuguese and English Wikipedias, we found that there is a noticeable activity increase during 2011, which, if we inspect closely, did occured in the Portuguese article.

IV. User Tests

User tests were carried out with 20 users: 16 male, 4 female, with an average age of 26 years, s = 7.8; 8 high school seniors, 9 bachelor degree holders and 3 master degree holders. Users were asked to execute tasks spread across several scenarios, which involved previously chosen articles. We focused, over 19 tasks, on verifying if they could understand the visualization and metrics, identify points of interest and patterns in the evolution of articles, and spot meaningful differences between different language versions of the same article. In short, 91.7% of users were able to spot activity rate changes, 80% managed to successfully compare and explain differences between articles, and 86% were able to use the visualization for information extraction tasks (finding answers/explanations to questions about the article's subjects). Alas, only 50% of users were able correctly obtain authorship information. The difficulty of this tasks was confirmed in a usability questionnaire, where users also pointed out that the system becomes harder to use for more information-dense articles.

V. Conclusions

Driven by the belief that Wikipedia's openness translates into a convergence of the trends and behaviors of its users towards public opinion, we decided to conceive a new visualization that shows these trends allowing the direct comparison of the evolution of different-language versions of the same article through time. We then analyzed several case studies, where we could find patterns, relate changes to real-world events and compare articles on different topics. User tests showed that while our visualization effectively enables users to identify activity patterns and compare articles, some more active articles make it harder for users to spot activity changes. In the future it will be interesting to extend the visualization with new metrics better highlighting the differences between the different article versions.

Acknowledgments

This work was supported by national funds through FCT Fundação para a Ciência e a Tecnologia, under project Educare - PTDC/EIA-EIA/110058/2009 and INESC-ID multianual funding - PEst-OE/EEI/LA0021/2013.

References

[1] Bao, P, et al.. Omnipedia: Bridging the Wikipedia language gap. In Proc, of the 2012 ACM CHI '12, pp. 1075–1084, New York, NY, USA, 2012. ACM Press.

[2] Brandes, U. and Lerner, J. Visual analysis of controversy in user-generated encyclopedias. In Proc. VAST 2007. pages 179 –186, Nov 2007. IEEE

[3] Chevalier, F, Huot, S. and Fekete, J.-D. Wikipediaviz: Conveying article quality for casual wikipedia readers. In Proc. PacificVis 2010 IEEE, pages 49 –56, March 2010.

[4] Mola-Velasco, S.M. Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals - Lab Report for PAN at CLEF 2010. In Notebook Papers of CLEF 2010 LABs and Workshops, Padua, Italy, September 2010.

[5] Suh, B., Chi, E., et al. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. In Proc. CHI '08, pages 1037–1040, New York, NY, USA, 2008. ACM Press.

[6] Viégas, F., Wattenberg, M., and Dave, K. Studying Cooperation and Conflict between Authors with history flow Visualizations. In Proc. CHI 2004, April 2004. ACM Press.