INSTITUTO SUPERIOR TÉCNICO
Universidade Técnica de Lisboa

PopCulture — Wikipedia as a mirror of society

Nuno João da Silva Garcia

Dissertation for the Degree of Master of Science in

Information Systems and Computer Engineering

Jury

Chairman:	Prof. Joaquim Armando Pires Jorge
Adviser:	Prof. Daniel Jorge Viegas Gonçalves
Member:	Prof. Rui Filipe Fernandes Prada

August 2012

Acknowledgements

Several people have helped and supported me during the past months. To those, family and friends, thank you!

I would also like to express my gratitude to Prof. Daniel Gonçalves, my adviser, for his endless support and guidance, and also for having managed to inspire and motivate me on every meeting we had.

Special thanks are in place to Felix Glaser and Pedro Gomes for their proofreading of this dissertation and of its extended abstract.

Nuno Silva
Lisbon, Portugal

Abstract

Wikipedia is a collaborative website which has been growing in popularity. As its popularity grows, so does the likelihood of it reflecting the opinion of society as a whole. For example, if an article has an unusually higher number of edits, we can conclude that it is associated with a popular topic. If an article is targeted by an unusual amount of edits made with the sole purpose of disturbing the work of other editors, this allows us to conclude that the article is about a controversial topic.

Thus, it is of interest to analyze Wikipedia data in order to draw conjectures about the public opinion, by condensing content, metrics and other information, along with their evolution through time, in a single visualization. Due to the significant amount of information involved, this is a goal other works have not excelled at, and which we try to accomplish with the work we present here.

After identifying a meaningful set of metrics, we crafted a unifying visualization that allows the easy analysis of the evolution of particular articles. Furthermore, several articles (even from different language-specific Wikipedias) can be directly compared in the same visualization, leading to additional insights on possible correlations.

Our user evaluation showed that the system has good usability, and that users are able to find meaningful insights. A series of case studies revealed that they could find patterns and infer hypotheses on the activity of Wikipedia articles. This is indicative of the adequateness of our solution.

Keywords: Wikipedia, User-generated content, Information visualization, Revision history

Resumo

A Wikipedia é um website colaborativo, que tem vindo a crescer em popularidade. Com este crescimento, aumenta também a probabilidade de a Wikipedia reflectir a opinião do público em geral. Por exemplo, se um artigo é alvo de um número invulgarmente alto de alterações, podemos concluir que está associado a um tópico popular. Se um artigo for alvo de um volume anormal de edições feitas com o único propósito de perturbar o trabalho dos outros editores, tal permite-nos concluir que o artigo aborda um tópico controverso.

É, por isso, interessante proceder à análise de dados provenientes da Wikipedia de forma a permitir a extracção de conjecturas respeitantes à opinião pública, condensando conteúdo, métricas e outra informação, tal como a sua evolução temporal, numa única visualização. Devido ao grande volume de informação manipulado, este é um objectivo que outros trabalhos não conseguiram atingir na perfeição, e que se tenta alcançar no presente trabalho.

Depois de identificado um conjunto de métricas relevantes, concebemos uma visualização que as unifica, permitindo uma fácil análise da evolução de artigos específicos. É, ainda, possível comparar directamente vários artigos (mesmo sendo de Wikipedias de línguas diferentes) na mesma visualização, levando a conclusões e correlações adicionais.

Os testes com utilizadores mostraram que o sistema apresenta uma boa usabilidade, permitindo aos utilizadores extrair conclusões relevantes. Um conjunto de casos de estudo revelou que os utilizadores conseguiram detectar padrões e extrair hipóteses sobre a actividade de artigos da Wikipedia. Isto é indicativo da adequabilidade da nossa solução.

Palavras-chave: Wikipedia, Conteúdo gerado por utilizadores, Visualização de informação, Histórico de revisões

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Traditionally, web sites were built on an "author hosts, guest reads" approach, where readers are not encouraged to provide their input, as that would have to be done through other media (such as electronic mail or Internet Relay Chat). This is cumbersome for the reader and also depends on the webmaster's will, who may simply ignore the feedback.

It was not until 1995 that Ward Cunningham suggested a new approach, where everyone is able to edit the hosted content[14]. This approach, which became known as "wiki", integrates feedback on the site itself, through the very same interface they use to read the content, with no need to use additional tools or wait for some review.

Users just have to follow an "Edit" hyperlink, which leads users to a special page where they can change the current content. When users are done with the changes, they just have to submit them, through a traditional web form, and the changes become immediately visible in the main version.

Most sites still follow the traditional approach, but the possibilities that arise from adopting this new approach led to several software packages and websites offering those new features, effectively empowering their readers with a power comparable to that of the original author.

The extent to which all these features are provided in sites following the new approach varies: some administrators may choose to provide only the quick feedback tool, employing a moderation process where changes are accepted or refused (examples of similar approaches are weblogs with moderated comments and protected pages in Wikipedia), possibly requiring a minimal "trust level" in order to bypass moderation (a scenario which is found, for example, in some of the Wikipedia protection levels[42] and in the Stack Exchange Network[1]). When all those features are provided together, administrators free their sites of the possible inconveniences of the traditional approach, especially the waiting time, which simply disappears.

Wikipedia, and other sites built upon the same MediaWiki software which powers Wikipedia, are implementations of this new concept, of which Wikipedia itself is frequently mentioned as an example. The name "Wikipedias" is also used as a way to refer to the whole family of wiki-like websites, which spans several areas, including serious encyclopedic content (Wikipedia itself[2]), news

[1] `http://stackexchange.com/`
[2] `http://www.wikipedia.org/`

(Wikinews[3]), specific-purpose encyclopedias on television shows, such as the "South Park Archives"[4] and the "X-Files Wiki"[5], or on books, such as the "One Wiki to Rule Them All"[6], a wiki site on the "Lord of The Rings" books by J. R. R. Tolkien, or even focused on jocular, sarcastic and ironic content, such as Uncyclopedia[7] (which, like Wikipedia, has localized versions, including Desciclopédia[8]).

Several software projects provide wiki-based sites to develop and hold either the main documentation or additional tips and guides, including Arch Linux[9] and Gentoo Linux[10]. There are also documentation wikis run by third-party administrators: for example, Alex Schroeder runs "EmacsWiki"[11], a wiki site devoted to documentation on GNU Emacs and XEmacs text editors, along with the Emacs Lisp programming language.

1.1 Motivation

Due to their user-centered nature, data available from wiki sites enables analysts to study the evolution of a page over time, assess the introduced changes and extract insights on the topic: trends, behaviors and their evolution over time. It also allows us to assess the opinions and interests of the site editors.

For example, by retrieving data on the Portuguese, English and Hungarian Wikipedia articles about the former Hungarian president Schmitt Pál, one can find out that the popularity of the article grew both in the English and Hungarian Wikipedias, probably as a result of the allegations of academic plagiarism involving his doctoral thesis[7], denoting a similar interest in the topic from both communities of editors.

On the other hand, by comparing with the Portuguese article, where the allegations did not spur a higher activity level, it is possible to conclude that the topic was not popular among Portuguese-speaking wikipedians.

As the popularity of Wikipedia grows, this possibility becomes even more appealing, as the assessed trends and behaviors may be close to the public opinion. For example, by comparing the Hungarian and Portuguese articles on Schmitt Pál, we may hypothesize that the accusations of plagiarism were not a popular or controversial topic for the population of Portuguese-speaking countries.

1.2 Wikipedia and MediaWiki

MediaWiki, the software behind Wikipedia, like other similar solutions, not only provides an editable site: unlike the initial version on Cunningham's wiki, it also keeps track of older versions of its pages.

Wiki platforms deal with older versions in different ways: Cunningham's system did not provide any mechanism at all, then introduced "Edit Copy"[2], a simple backup mechanism that allowed users to

[3] http://www.wikinews.org/
[4] http://southpark.wikia.com/
[5] http://x-files.wikia.com/
[6] http://lotr.wikia.com/
[7] http://uncyclopedia.wikia.com/
[8] http://desciclopedia.ws/
[9] https://wiki.archlinux.org/
[10] http://wiki.gentoo.org/
[11] http://www.emacswiki.org/

keep a copy of the previous version, and was later extended to offer "History Pages"[1], which, albeit still time-limited, stores several older versions for the same page.

On the other hand, MediaWiki stores all versions on its database, with no time restriction. Older versions, in general, allow collaborators to undo unconstructive or undesirable editions; In this case, it also eases the task of analyzing the trends of an article, as the revision history is available through the MediaWiki API[37].

1.3 Goal

Our goal is to *use information visualization techniques to enable the analysis and assessment of trends in Wikipedia, through the presentation of metrics and other information on (or from) the revision history of Wikipedia articles.*

With this, we aim at enabling the user to easily extract novel insights and conclusions, regarding the public opinion.

We also want to allow 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[12].

1.4 Contributions

Our work introduces a new visualization system for Wikipedia articles and their history, which focuses on the consolidated display of the sequence of metrics computed for the article revisions. We also present a novel set of metrics, loosely based on insights from related works, which is then used in our visualization.

The main contribution lies, perhaps, in the way our system aggregates informations from the entire revision history of a Wikipedia article, effectively allowing users to spot the overall trend of the article from a single screen. Together with the cached, integrated display of article content, the system enables users to explore the evolution of the article in a way that would not be possible through the plain Wikipedia interface.

We also contribute the metrics component: a library which is able to compute these metrics for each revision of a Wikipedia article, on top of which other systems may be built, even systems other than visualizations.

We chose not to rely on any specific wiki site, but, rather, on the MediaWiki API. As the popularity of the wiki approach increases, we expect more sites to adopt this approach, some of which will choose the

[12] Although, in many cases, it will be possible to infer conclusions on specific countries or geographical areas, it is important to highlight the distinction between "country" and "language", as the several editions of Wikipedia operated by the Wikimedia Foundation are *language-specific*, not country-specific. This becomes even more important when the language of an edition of Wikipedia is geographically widespread, such as with Portuguese, which is an official language in America (Brasil), Africa (Angola, Cabo Verde, Guiné-Bissau, Guiné Equatorial, Moçambique, São Tomé e Príncipe), Asia (Macau (China), Timor-Leste) and Europe (Portugal), even if it is only spoken by more than 80% of the population in Angola, Brasil, Portugal and São Tomé e Príncipe[40].

MediaWiki package, effectively enabling our system to analyze articles from those sites and compare articles from completely unrelated sites.

1.5 Document Structure

In chapter 2, we start by surveying related works on visualizations, either Wikipedia-specific or broader attempts at the graphical display of information and patterns. We then complement this analysis by surveying previous works on the analysis of data from Wikipedia.

After describing those works, we then collect the heuristics and informations used by their authors and evaluate them, proposing initial guidelines on the algorithms and design choices that will drive the design of our visualization, leading to the data analysis component of the system, which is described in detail in chapter 3.

The same analysis is done on the works on visualizations, leading to our choices concerning the visualization part of our system, presented in chapter 4, where we also illustrate its applications through several case studies.

After presenting our system, we proceed with its evaluation with end-users, a process described in detail, along with its results and conclusions, in chapter 5.

Finally, we conclude this work with final remarks on its result (chapter 6).

We also provide a glossary of Wikipedia concepts in appendix A.

Chapter 2

Related work

Wikipedia has already been the focus of several academic works, which we survey in this chapter, in order to summarize and discuss the state of the art. We start with works concerned with the visualization of Wikipedia data and metrics, in section 2.1. Then, in section 2.2, we survey works focused on the direct analysis of Wikipedia data, which do not involve visualizations.

2.1 Visualization

Several approaches have been previously taken to visualize Wikipedia-related data or other multi-topic corpora. We first analyze some of these works in the sections which follow, and then their characteristics are summarized and discussed in section 2.1.13, leading to several remarks on our own work.

2.1.1 WikiViz

When working at AT&T Labs, Chris Harrison started conceiving a visualization tool that could be used with Wikipedia to represent connections between topics, highlighting organization and making it possible to find unexpected connections.

This led to WikiViz[18], a mindmap-like visualization of Wikipedia categories and their connections, initially done with a spring model (so that related nodes get rendered closer to each other, and unrelated nodes are pulled apart) and with real-time rendering.

The tool was later changed to use another algorithm, as the spring model did not scale well, and to render the map to a file instead of relying on real-time on-screen rendering. With these changes, it became possible to render high-quality maps (an example of which is shown in figure 2.1), instead of having to comply with the low-quality resolution of a typical computer screen.

While WikiViz allows the easy presentation, in an image, of the importance and relations of the topics, it does not provide room to convey additional information and does not have a way to depict the evolution through time.

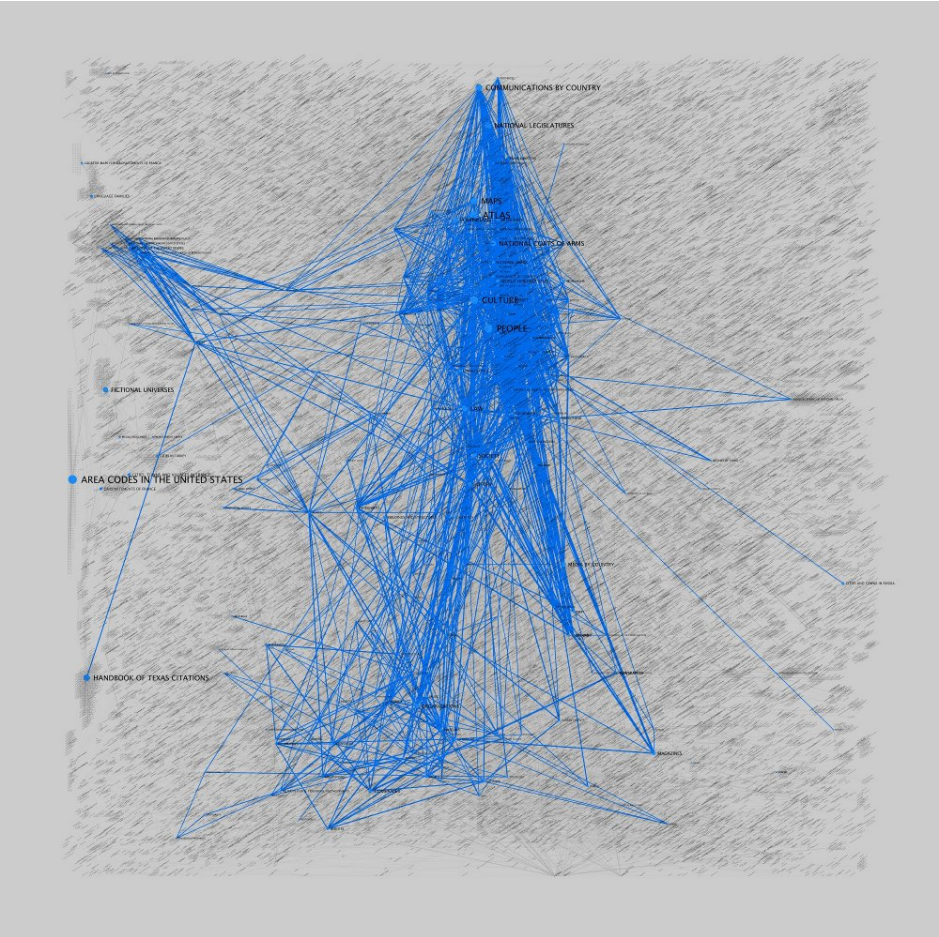


Figure 2.1: WikiViz visualization for "Politics" (whose colors have been inverted).

2.1.2 ClusterBall

In another project, Chris Harrison conceived ClusterBall[17], which, like WikiViz[18], is a mindmap-like visualization that represents category pages and links between them, but which bounds the visualization to a ring, with nodes either inside or on the ring (figure 2.2).

The ring is centered on a main, central node, and has three visualization levels:

- The central node itself;
- Directly linked pages, which are drawn inside the ring, between the central node and the ring boundary;
- Pages linked to the directly linked pages, which are drawn *over* the ring.

Nodes for directly linked pages are adjusted so that it is possible to group some of them in clusters, enriching the visualization.

ClusterBall has the strengths of WikiViz, while making the relations and weights clearer. On the other hand, it also suffers from the same issues, making it hard to integrate additional information.

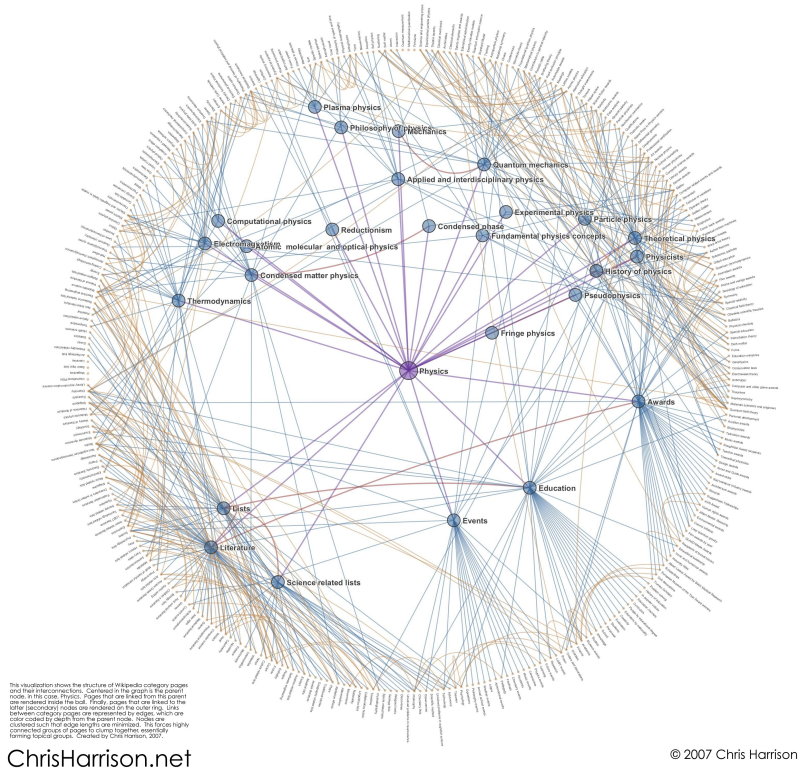


Figure 2.2: ClusterBall visualization for "Physics".

2.1.3 Wikiswarm

Jamie Wilkinson applied an existing collaboration visualization tool, `code_swarm`, to Wikipedia, creating a system to visualize collaborations and editions on a set of articles (or from a set of editors), Wikiswarm[46].

`code_swarm` was conceived by Michael Ogawa, as part of his (visual) study of communication and collaboration in software projects, and departs from the more traditional approach of assigning data to a specific point in space, relying instead on a representation of entities as moving elements, as an "organic information visualization".

This results in a video-based visualization, which depicts the evolution of communication and collaboration over time, highlighting how much articles did a contributor edit and how recent are those editions, as shown in figure 2.3.

While the visualization does not show more than authors and relationships among them, it does provide a clear way to follow changes on these authors and relationships through time.

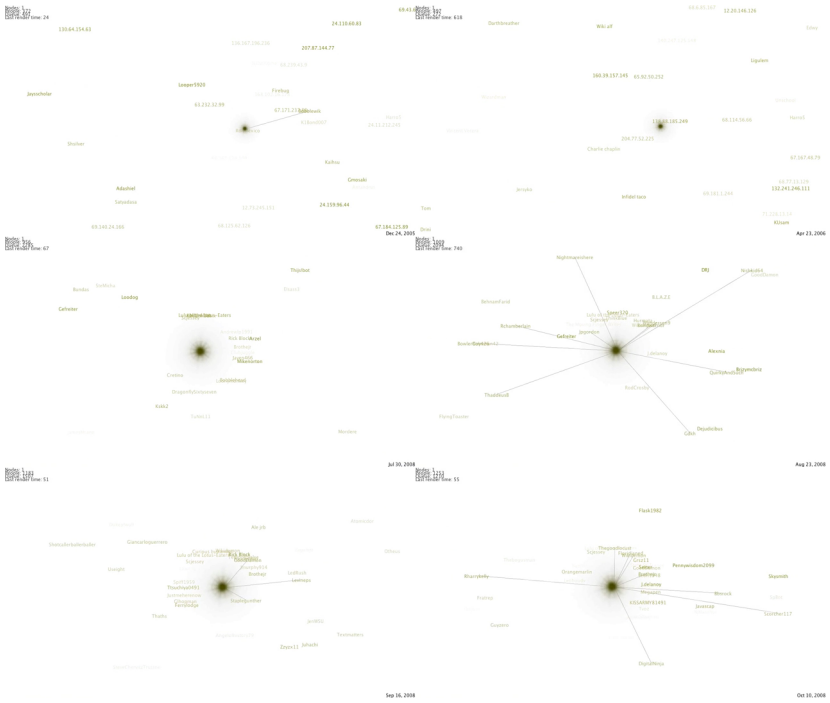


Figure 2.3: Frames from the Wikiswarm movie visualization for the article on “Barack Obama”.

2.1.4 History flow

Viégas, Wattenberg and Dave[32, 31] investigated how Wikipedia can work while being open and vulnerable. To analyze Wikipedia data, they developed a new visualization method, “history flow”, where a version is represented by a vertical “revision line” (figure 2.4), whose length is proportional to the length of its text.

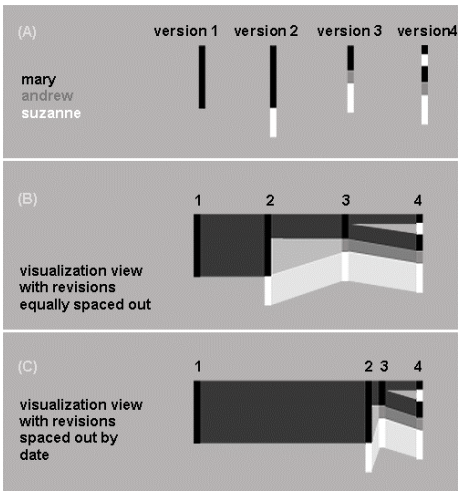


Figure 2.4: History flow “revision line”, with examples of color coding, equal spacing and time-based spacing.

Sections of a “revision line” are colored according to their author (each editor is assigned a different color). Sections of text that have been kept between consecutive editions are linked by drawing *shaded*

connections between these segments. This also highlights insertions and deletions, which appear as a gap among the shadows.

Their visualization tool offers the ability to filter editions by editor, highlighting only her editions. The tool also allows the user to select a location on the graph — the tool then shows the text for the selected position of the selected edition. When the user selects a revision line, the editor's comment for that edition and its date are shown and all other editions by the same author are highlighted.

The Wikipedia text matching algorithm works at the paragraph level, but the authors decided to use another algorithm, based on a smaller unit, sentences, as it performs better.

Revision lines can be equally spaced or spaced proportionally to the time between editions. Due to the gaps in the visualization caused by insertion and deletion, when revision lines are evenly spaced, some kinds of vandalism are easy to spot in their tool: mass deletions will result in a completely empty column, in contrast with the other editions. On the other hand, spacing revision lines by date hides these empty spaces, highlighting how quickly most of the vandalism gets fixed (figure 2.5).

"Edit wars" can also be spotted, as these lead to a "zigzag arrangement" along several editions (figure 2.6).

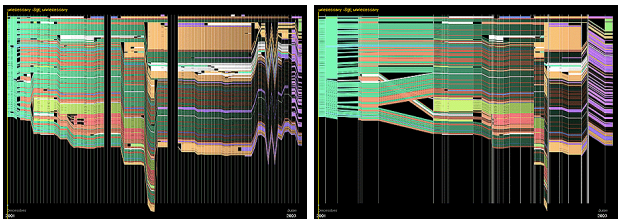


Figure 2.5: History flow visualizations of the "Abortion" article: the left one is evenly spaced, while in the right one revisions are spaced by date, highlighting how quickly vandalism gets fixed in Wikipedia.

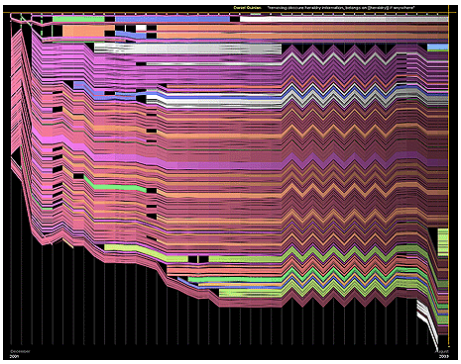


Figure 2.6: History flow visualization of the "Chocolate" article, showing a "zigzag arrangement".

This visualization allowed the authors to spot other patterns: insertions and deletions are more frequent than text moves and the text from the first version has the highest "survival rate". It is also observed that the length of a page does not tend to stabilize over time.

Suggested future work includes extending the statistical analysis used to corroborate the observed patterns and relations; exploiting the relationship between talk and article pages to find out how talk page discussions compare to the observed collaboration patterns; and compare the English Wikipedia with other language Wikipedias and other (non-Wikipedia) wiki sites.

2.1.5 ThemeRiver

Havre, Hetzler, Whitney and Nowell[19] proposed a river metaphor to represent theme changes over time. The river flows left-to-right, representing time flow; variations in the importance of a theme are represented by changes in the river width; and different themes are represented by differently color-coded segments of the river (figure 2.7).

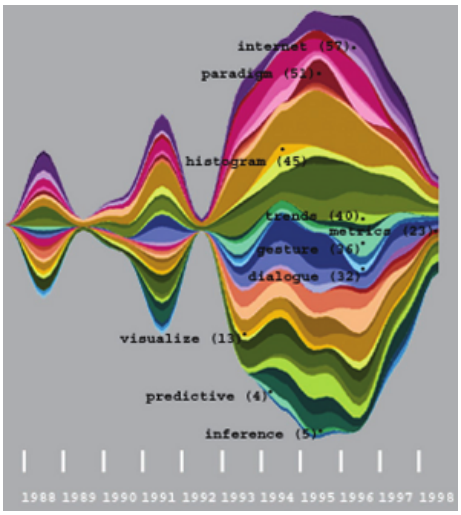


Figure 2.7: A ThemeRiver visualization.

Related themes are grouped and colored using a "color family", to highlight their changes.

This view requires themes and their importance to be assessed beforehand, and is expected to represent only a subset of the themes present in the documents analyzed. In order to allow users to check whether a narrow river is due to lighter documents or to a change in themes, a small histogram, representing the amount of documents analyzed and the documents containing the represented themes, can be added to the graph.

External events, from a given list, can also be flagged in the graph, to help users identify possible patterns or connections (figure 2.8).

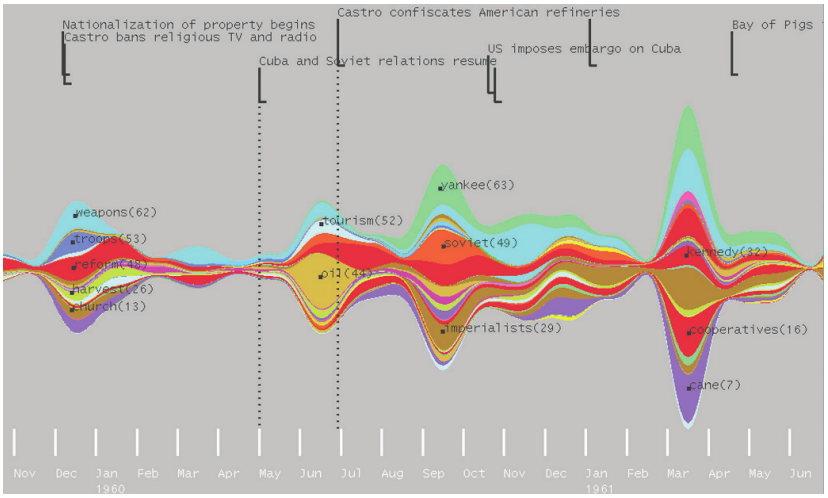


Figure 2.8: A ThemeRiver visualization, with external events flagged above the graph.

This view allows users to distinguish persistent themes from bursty ones, to detect subtle changes and to identify macro trends. During the usability evaluation, users also highlighted that the connectedness of themes in the river representation made it easier to follow a trend over time, compared to a histogram.

The authors identified some challenges with this visualization: the number of themes that can be represented is limited, which raises scalability issues. Possible suggestions include representing *sets of themes* instead or relying on *color families* to make room for more themes. Another scalability issue lies in the computational costs: faster algorithms are desirable, in order to allow for a more interactive visualization.

Another issue lies with the "truthfulness" of the graph at a certain level of detail: lines are drawn using interpolation, which is just an approximation.

Some of the suggestions made by users during the usability evaluation remain as future work, such as the ability to reorder theme currents.

2.1.6 Revert Graph

Kittur et al.[30] developed Revert Graph, a conflict pattern visualization tool that relies on a model based on edit histories and relation between edits, with emphasis on "reverts" (figure 2.9).

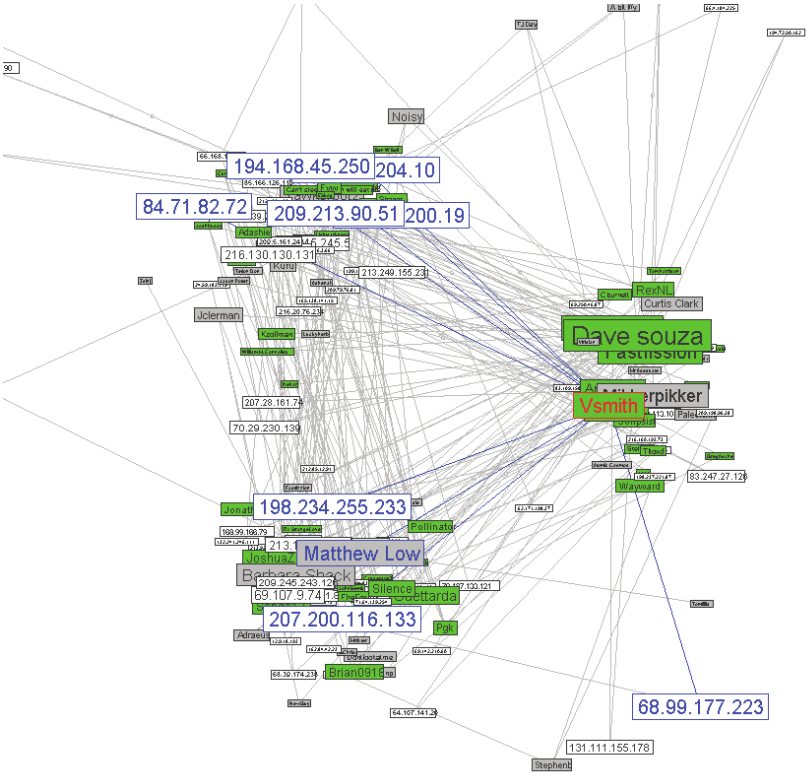


Figure 2.9: Revert Graph visualization for the "Charles Darwin" article.

Due to the size of the Wikipedia data set, which makes an exact disagreement model infeasible, the authors decided to *approximate* disagreements, using *revert revisions* to detect such conflicts, as that

kind of edition is seen by editors as a sign that some editor strongly disagrees with some previous edition.

Their tool detects reverts in two ways: MD5 hashes are generated for every edition, and when a hash from some later edition matches the hash for a previous one, that later edition is considered a revert; they also look for user-labeled reverts (which include "revert" or "rv" in the edit comment).

The proposed model has some additional properties, in order to address some challenges with conflict models based on reverts: self-reverts are ignored; the degree of conflict between two users is defined (as the number of reverts between these users); two users who only make reverts against the same third user and do not make revert edits against each other are said to have similar opinions; and reverts which span several editions are only counted as reverts against the last edition.

In order to highlight the social dynamics captured by this model, the visualization groups users with compatible opinions together, and separates users with incompatible opinions. This was implemented using a *force-directed* graph layout, where nodes (users) are comparable to particles, with gravitational fields, and edges (reverts) are comparable to springs. That is, users are grouped together unless there is some revert relationship between them.

Nodes are also formatted in a way that makes it easy to understand the role and impact of the user: the size of editor nodes is proportional to the logarithm of the number of reverts they made. Nodes are also color-coded: administrators are green, normal and registered users are gray, and unregistered anonymous users are white.

The authors identified several patterns using this visualization, after having also browsed additional information on the edits and the involved users, to get a clear insight on the involved opinions.

- **Clusters and opinions:** The authors confirmed that the user groups identified by this visualization correspond to distinct opinion groups and that these groups also correspond to major points of view on the discussed topic.
- **Mediation:** Some groups of users revert edits from many other groups, a signal that they are mediating conflicts and moderating the discussion, in order to achieve a neutral article.
- **Fighting vandalism:** Disputes based on vandalism are easily identifiable, as they usually occur between unregistered users (white) and administrators (green), so the graph on vandalism fighting would have a group with a large amount of white nodes with revert relationships connecting them to a group with a large amount of green nodes.
- **Controversial editor:** editors who made a lot of reverts are highlighted, being shown with a larger node size. This allows the easy identification of highly controversial users.

The authors conclude that this visualization was able to highlight unexpected controversies that would probably be hard to find without the visualization. They also recognize some limitations with the visualization: the force-directed layout may not create optimal user groups if there are not enough revert relationships. Due to the nature of the model, it is also not possible to detect conflicts between editors who did not revert each other's edits.

As the tool does not cover all details of social dynamics, the authors mention possible future work on this area, by answering some questions like what types of conflicts are there, which are the possible sources of disagreement and what is the motivation behind editing.

2.1.7 Visual Analysis of Controversy in User-generated Encyclopedias

Brandes and Lerner[9] developed a network visualization of editions (figure 2.10), where users who edit content are linked to the users whose editions they change, through weighted edges (weights are a measure of how quick the change was made[1]). This is based on the fact people are more likely to respond to a message if they disagree with it. The authors also suggest increasing the weight if the edition is a revert.

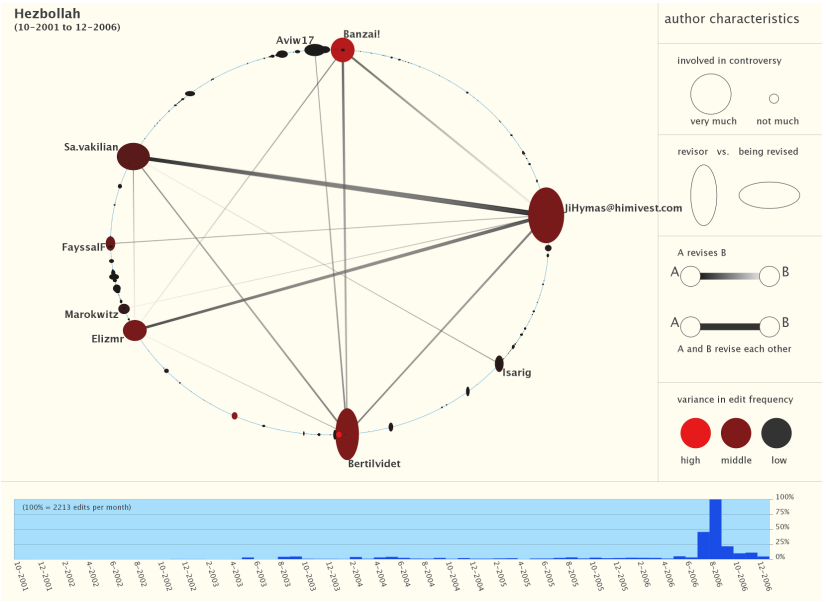


Figure 2.10: Brandes and Lerner edition network visualization for the "Hezbollah" page. The bar chart shows a higher edit volume during July and August 2006, when the Israel-Hezbollah War took place.

Editors who change each other editions frequently are shown in opposite sides of the image, and independent conflicts do not overlap (if there are two independent conflicts, one will be represented horizontally and the other vertically, and so on).

To avoid having some editors placed near the center of the visualization, positions are normalized on an ellipse, so that all editors are over the ellipse.

Editors are also drawn as ellipses, whose width-to-height ratio is proportional to the ratio between the weight of outgoing edges and the weight of incoming edges[2].

Edges between editors who revise and editors who are revised are filled with a dark-gray to light-gray gradient. (If the editors do not have distinct roles, the edge is filled with a uniform dark gray.) Edge tickness is proportional to edge weight, and the visualization only shows the highest weighted edges.

[1] For several changes, weights are summed up
[2] Ratio between their "degree as a revisor" and their "degree of being revised"[9].

To help telling editors who are interested in the page from editors who make several edits but then become uninterested, or whose interest is unstable in some other way, the former editors are colored black and the latter are colored red.

The visualization also includes a bar chart with the edit intensity over time, which helps finding out when was the article a "hot topic" and also other interesting time periods to focus on (the visualization can be restricted to certain time intervals). The tool also separates users in sub-networks that can be analyzed separately, in order to tell apart unrelated disputes.

The authors propose future work in the analysis of the influence of news events on some articles, to find out how does that influence change controversy; authors also propose telling opinion disputes from vandalism disputes, and considering the authorship of changed (revised) text when building the revision network.

Although their visualization does not convey more than relationships between users, it provides a plot showing the article trends along time. The visualization lacks an integrated way to convey evolution along time, but the time range restriction, together with the plot, enables users to browse through visualizations of controversy along time.

2.1.8 Analyzing and Visualizing the Semantic Coverage of Wikipedia and Its Authors

Holloway, Bozicevic and Börner[20] proposed a semantic map visualization of Wikipedia categories, using color-coding to identify major topic areas (figure 2.11 left), and to tell recent editions apart from old editions (figure 2.11 middle). Two categories are considered similar if they occur together in the same article.

After computing the similarity values (used to weight edges in a graph), the final image is rendered using a separate tool, Pajek.

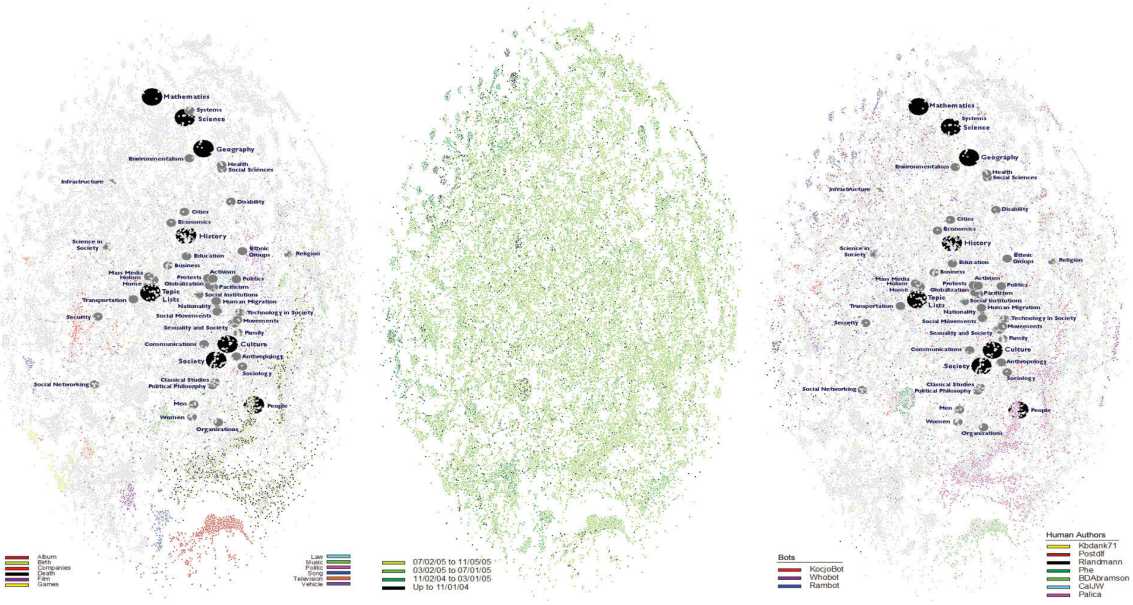


Figure 2.11: Holloway et al. visualization of the English Wikipedia category network, color-coded by category (left), by edit age (middle) and by most active editors (right).

The most active editors are also highlighted in the visualization, to help assessing topic coverage on an editor-by-editor basis (figure 2.11 right).

Like WikiViz, this visualization, while picturing the whole semantic network on a single screen, does not convey additional information and, while it allows users to highlight topics by their recent activity rate, it does not allow the analysis of the semantic network through time.

Suggested future work includes studying the influence of categories in the growth of Wikipedia, and on extending the visualization to articles. The authors also propose examining non-English Wikipedias, to assess differences and similarities.

2.1.9 WikiDashboard

Suh, Chi, Kittur and Pendleton[29] presented a tool, WikiDashboard[3], which embeds a dashboard in Wikipedia pages, providing an overlay with visualizations about the current article (or user, if the article is a user page).

In articles (figure 2.12), the dashboard shows an edit activity graph, which shows the weekly activity of the article, in a gray line, and of the associated talk page, in a blue line. This graph is followed by the list of the top active editors (for that article) ordered by their edit activity, which is also highlighted in a graph for each editor.

In user pages (figure 2.13), the dashboard displays the user editing patterns, with the top graph showing the user's weekly activity, followed by the pages where the user has made more edits.

User names and article titles are hyperlinks which allow the user to quickly browse their pages.

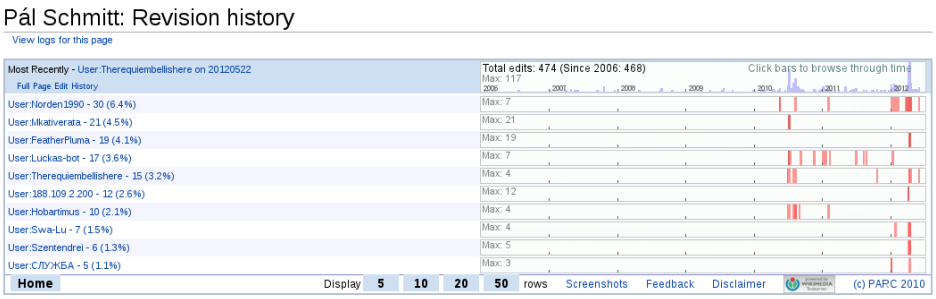


Figure 2.12: WikiDashboard visualization for the article on Pál Schmitt, former Hungarian president, showing a peak in the first half of 2012, possibly coincident with the date when he was stripped of his doctoral degree.

[3]Currently available at `http://wikidashboard.appspot.com`.

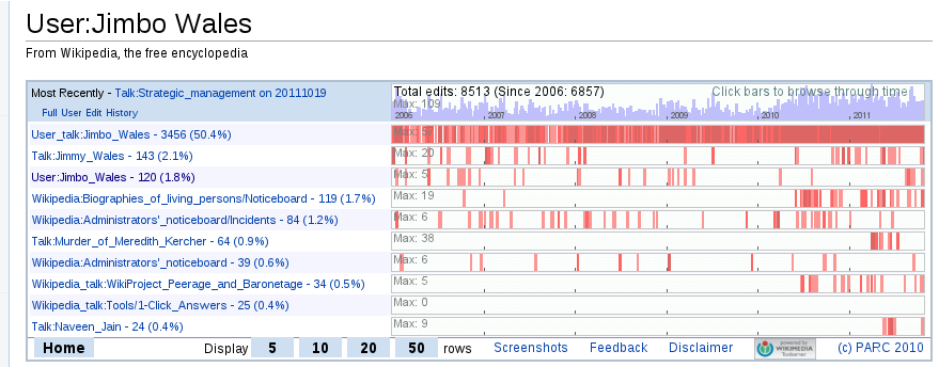


Figure 2.13: WikiDashboard visualization for Jimbo Wales' Wikipedia user page.

The authors mentioned possible future work, including enriching the dashboard by highlighting some information, like the most recently active editors, in order to provide richer context, and accounting for the weighting of bigger edits compared to minor fixes, possibly using the number of words changed as a metric. They also made a remark on the possible impact of people who "game the system" to get in the top 10 presented by WikiDashboard.

2.1.10 WikipediaViz

Chevalier, Hout and Fekete[11] collected several heuristics in their visualization system, WikipediaViz. Their goal was to provide enough information to help the user assess the quality of an article at a glance.

First, they identified the heuristics employed by expert Wikipedia users when assessing article quality:

- "Word count" — quality articles are long (although long articles are not necessarily good);
- "Number of Contributors", "Rate of Contribution" — allows to assess the number of distinct major contributors (of which high-quality articles have a larger number);
- "Number and Lengths of Edits" — allows to identify incomplete articles, articles on controversial topics and vandalized articles;
- "Number of References and Internal Links" — identifies underlinked and overlinked articles, and articles that lack enough references to back their content;
- "Length and Activity of the discussion" — highlights possible controversies, either past ones and current ones.

Five visualizations, based on the metrics above, were included in a modified Wikipedia interface:

- "Word Count" — shows the number of words in the article;
- "Authors Contributions" (figure 2.14) — shows a pie chart where each editor (author) has a slice whose area is proportional to the amount text they contributed that is still present in the current edition of the article);

- "Article Timeline" (figure 2.15) — shows a graph of article length (bar height) over time. The bar color is used to encode activity (darker colors correspond to more intense activity). Small icons (figure 2.16) are shown to signal important events like banners (red flags) and protections (lock icon);
- "Internal Links Meter" (figure 2.17) — displays, in a gauge-like visualization, the link density of the article, also allowing the user to easily spot under- and overlinked articles;
- "Discussion Length and Activity Indicator" — shows the length of the associated talk page (if it exists). If there has been recent activity there, it also shows a link to the page.

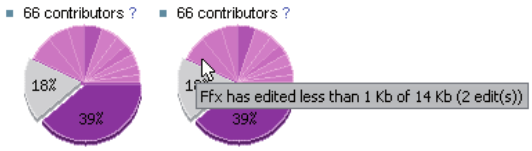


Figure 2.14: WikipediaViz "Authors Contributions" visualization.

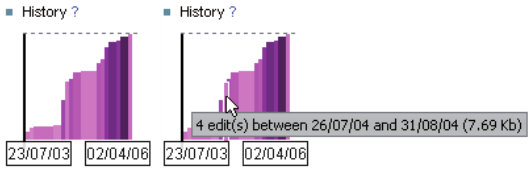


Figure 2.15: WikipediaViz "Article Timeline" visualization.

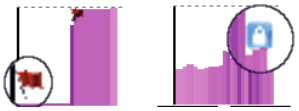


Figure 2.16: WikipediaViz "Article Timeline" visualization: flag and lock icons.

Figure 2.17: WikipediaViz "Internal Links Meter" visualization.

Studies carried with users concluded that using WikipediaViz does not cause loss of precision in quality assessment, and that it speeds up the assessment process.

The authors proposed future work involving the implementation of a live WikipediaViz system and the execution of more studies with users, to better assess the benefit provided by the system.

2.1.11 iChase

Richie, Lee and Chevalier[26] presented an edition history visualization, which comprises two heatmaps over a timeline (figure 2.18).

Figure 2.18: iChase visualization of WikiProject "Louvre Paintings".

As the proposed visualization displays articles, contributors and editions, these are color-coded with different colors (green, blue and purple, respectively).

The visualization has several main components:

Timeline (figure 2.19, (1) and (2)) represents the flow of time in the visualization, and can be used to adjust the granularity and the date range.

Activity heatmaps (figure 2.19, (3)) are maps between entities (articles or contributors) and time periods, whose cells are color-coded according to the number of editions.

Rendering is not done at cell level, but instead at the smallest level possible, with the trend corresponding to that cell being assessed from its overall color pattern.

Heatmaps can be scrolled and its rows and columns can be collapsed.

Aggregated row indicators (figure 2.19, (4)) convey metrics about the entire row — for articles, the number of editions and the number of active editors. Information is shown in color intensities, but also in numeric form, to ease comparison.

Activity line graphs (figure 2.19, (6)) show the evolution over time of the number of edited articles and the number of active editors.

Legend (figure 2.19, (5)) provides summarized metrics about the whole visualization: the total numbers of articles, editors and editions shown in the visualization.

Figure 2.19: iChase visualization of WikiProject "French Revolution", with the several visualization elements numbered.

Tooltips with details are also provided when the pointer is hovered over a cell (figure 2.19, (9)). The interface also provides links to Wikipedia, in order to show differences between editions. In the heatmaps, cells and columns can be expanded to be examined with more detail (figure 2.19, (7) and (8)).

A user study was conducted, where Wikipedia administrators stated that it was easy to spot vandalism in iChase, and that this visualization makes it easier to find new contributors and discover their interests. Administrators also rated iChase better than the other tools they currently use (either Wikipedia's watchlist or LiveRC, a live, real-time version of the list of recent changes).

Planned future work includes carrying out more user studies, fixing identified usability problems (regarding the way columns and rows are closed), and adding already identified missing features, such as differentiating the contributor according to their type (anonymous, robot, registered, administrator, ...), identify reverts as such in the visualization, links to perform tasks in Wikipedia (such as reverting editions and blocking contributors), color-code regular and registered contributors, the ability to filter out already reviewed articles and registered editors, and more sorting options.

2.1.12 Omnipedia

Bao, Hecht, Carton, Quaderi, Horn and Gergle[6], driven by the differences among Wikipedia editions, proposed a visualization, Omnipedia, that enables users to see the breadth of and differences in topic coverage across different editions of Wikipedia.

Their work aims to show the different, possibly culture-dependent, viewpoints, and to highlight isolated content which does not exist in the user's native language edition of Wikipedia.

Omnipedia works by focusing on a "multilingual article": when the user looks for a concept, a visualization is built based on the coverage of that concept across several editions of Wikipedia, relying on Wikipedia's own interlanguage links, and on a user-defined granularity setting, that deals with split articles[4], and employing heuristics and rules to handle ambiguities.

[4]In some languages, a topic may be discussed with more detail than in other languages, possibly rendering the article too big that it has to be split in several subarticles, while in the other languages it remains as a single article.

Circles are shown ordered and color-coded by language, with the topics only covered in one edition being stacked on the left, while topics with multilingual coverage are shown, as circles with more than one color, on the right side of the visualization (figure 2.20).

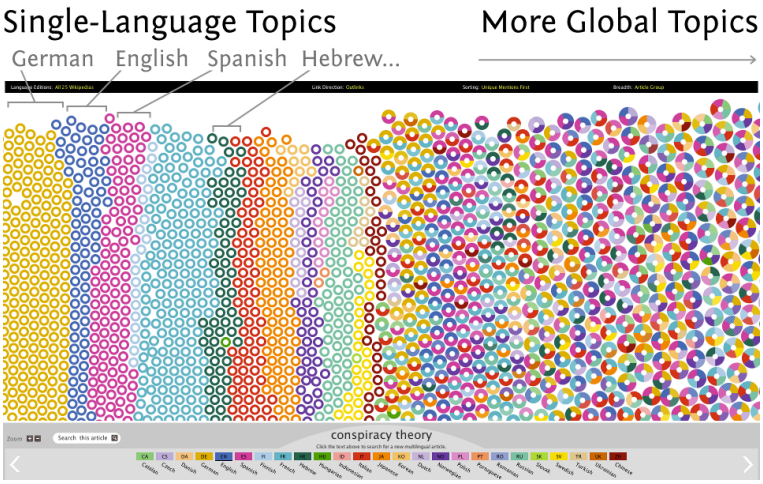


Figure 2.20: Omnipedia visualization, showing topics discussed in a single language (left) and in multiple languages (right).

Interaction is done in the "interface language", with topic names being translated if they are discussed in the interface language (otherwise the native name is used). Omnipedia supports 25 languages, from which the users can select a language subset of their own choice. Users can double-click the circles to navigate through the topics, while single-clicking highlights related topics and shows a machine-translated portion of text depicting the coverage of the selected topic (figure 2.21).

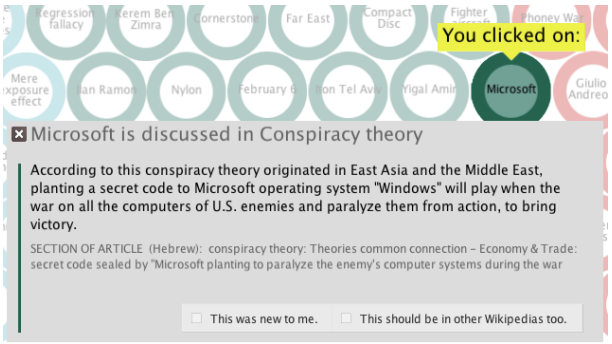


Figure 2.21: Omnipedia showing a small portion of machine-translated text on the currently selected topic.

Omnipedia was then tested with users, concluding that users took advantage of the visualization to find out which topics were widely discussed, across language boundaries, also using the system to spot differences in the coverage of those topics. Users also gained, through their use of the system, insight on how different can the coverage of topics be across several languages, expressing surprise and curiosity on topics only discussed in a single language other than their native one.

The authors propose future work regarding the application of the same visualization strategy to other media, such as Twitter and Flickr, which present similar linguistic barriers. As Omnipedia also relies on a new "relatedness" measure to build the visualization, the authors mention the possibility of enhancing the algorithm and applying it in other interlinguistic contexts.

2.1.13 Summary

The surveyed works can be grouped by the type of visualization they employ:

- **plot-based visualizations:** History flow[32, 31], ThemeRiver[19], WikiDashboard[29], WikipediaViz[11].

  These visualizations are the only, other than Omnipedia[6], that depict content. This may mean the other types of visualizations are unsuitable for content-based analysis.

- **graph-based visualizations:** Revert Graph[30], Brandes et al.[9], Holloway et al.[20], WikiViz[18], ClusterBall[17], Wikiswarm[46], Omnipedia[6].

  Of all visualizations, the only one which operates at the article content level is Omnipedia, which presents it through an additional, tooltip-based visualization. This suggests that graph-based representations may be less suitable to visualize information at this level.

- **heatmap-based visualizations:** iChase[26]. Although this visualization does not depict content nor controversy, it was the sole surveyed work presenting a heatmap-based visualization, what prompts us to be conservative with regard to conclusions on the unsuitability of this kind of visualization to incorporate those features.

Several features are present along all types of visualizations: *authorship*, visualization *along time* and the *author importance*, all of which are also used in more than half of the previous works. The most common type of visualization was the graph-based one, followed by plot-based visualizations.

We observed that, in all the surveyed works, authors chose just one kind of visualization. This is due to the fact two different and sophisticated visualizations cannot be easily joined.

The visualizations have different goals and, according to their goal, authors chose to focus on different levels,

- visualizations which focus on an **article**: History flow[32, 31], ThemeRiver[19], Revert Graph[30], Brandes et al.[9], WikiDashboard[29], WikipediaViz[11], Wikiswarm[46];
- visualizations which focus on a **network of articles**: Holloway et al.[20], iChase[26], WikiViz[18], ClusterBall[17], Omnipedia[6];

By analyzing the several visualizations, it was possible to devise the following feature set:

- **Content**, either the visualization of a graphical entity derived from content (History flow[32, 31], ThemeRiver[19]) or the presentation of Wikipedia content to the user (WikiDashboard[29], WikipediaViz[11], Omnipedia[6]);
- **Authorship**, regarding the inclusion of information that enables the user to identify different authors in the visualization (History flow[32, 31], Revert Graph[30], Brandes et al.[9], WikiDashboard[29], WikipediaViz[11], iChase[26], Wikiswarm[46]);
- **Along time**, the ability of the visualization to convey, either directly, or through an additional element, the chronological evolution of the depicted data (History flow[32, 31], ThemeRiver[19], Brandes et al.[9], Holloway et al.[20], WikiDashboard[29], WikipediaViz[11], iChase[26], Wikiswarm[46]);

- **Topic importance** is mostly common among visualizations that focus on networks of articles (Holloway et al.[20], iChase[26], WikiViz[18], Omnipedia[6]), where the strength of each article (a topic) is conveyed, but is also used by ThemeRiver[19], which splits a single source (an article) into several topics.
- **Author importance** regards the weight of the editions contributed by a specific author on the entire article (History flow[32, 31], Revert Graph[30], Brandes et al.[9], WikiDashboard[29], WikipediaViz[11], iChase[26], Wikiswarm[46]);
- **Controversy**, either the portrayal of controversy networks which show the controversy between different authors (also highlighting the article controversy level), as is the case with Revert Graph[30] and Brandes et al.[9], or a visualization presenting the article in such way that controversy becomes directly observable (History flow[32, 31]).

It is possible to identify a set of core features for plot-based representations: all of these depict content and topics in some way. While there is no common set of features in graph-based visualizations, these visualizations cover the entire feature spectrum.

The visualizations, along with their type, focus and features, are summarized in table 2.1,

	History flow[32, 31]	ThemeRiver[19]	Revert Graph[30]	Brandes et al.[9]	Holloway et al.[20]	WikiDashboard[29]	WikipediaViz[11]	iChase[26]	WikiViz[18]	ClusterBall[17]	Wikiswarm[46]	Omnipedia[6]
Type	P	P	G	G	G	P	P	H	G	G	G	G
Focus	A	A	A	A	N	A	A	N	N	N	A	N
Content	✓	✓				✓	✓					✓
Authorship	✓		✓	✓		✓	✓	✓			✓	
Along time	✓	✓		✓	✓	✓	✓	✓			✓	
Topic importance		✓			✓			✓	✓			✓
Author importance	✓		✓	✓		✓	✓	✓			✓	
Controversy	✓		✓	✓								

Type key: **P** – plot, **G** – graph, **H** – heatmap.
Focus key: **A** – Article, **N** – Network of articles

Table 2.1: Comparison of visualization techniques employed by previous works and their main features.

2.2 Data analysis

We will now survey several previous works focused on the analysis of Wikipedia, with no associated visualization system. First, we start by surveying, in the sections which follow, these works. Details are then summarized in section 2.2.7, where we present the implications for the design of our solution.

2.2.1 On Ranking Controversies in Wikipedia: Models and Evaluation

Vuong et al.[33] present several analysis models to automate the assessment of the controversy of articles, using metrics based on page history.

The authors define a metric to rank known controversial articles, the *Article Tag Count*, *ATC*, that ranks pages according to the number of dispute tags in all its editions. Unfortunately, as the authors warn, this only covers a small number of articles.

ATC is then used to evaluate two new models: the *Revision Count model*, which assumes that an article with more editions is more controversial, and the *Contributor Count model*, which assumes that an article with more editors is more controversial.

These models do not recognize disputes, can be easily abused and are unable to detect highly controversial articles[5].

Then several models are proposed. Each model defines the controversy of an article and the controversy of an editor:

In the **Basic Model**, the *controversy of an article* is computed from the impact of deletions on editions; the *controversy of an editor* is defined from the controversy of the articles where the editor added or removed content.

The **Controversy Rank Models** are based on the Basic Model, but their values are weighted by inverse controversy scores[6] (so dispute among aggressive, controversial editors does not affect score as much as disputes among non-aggressive, non-controversial editors).

As the controversy of the article is weighted by the controversy of two contributors (the one that deleted the content and the one that wrote the content), these two values must be combined using some function. According to the chosen function, the authors define two models: average ("CR Average") and product ("CR Product", on which disputes between high controversy and low controversy editors have a factor close to zero).

Article age is defined as the number of editions, and an age-dependent weighting function is defined, so that the weight is close to 0 for a young age, 0.5 for a median age, and close to 1 for an old age.

Using these functions, the authors then define age-aware variants of the Basic, CR Average and CR Product models, where the article controversy is weighed with the age-dependent weighting function (the editor controversy remains unchanged).

Evaluations of non-age-aware models and age-aware models led to the conclusion that the CR Product Model outperforms all others (both non-age-aware and age-aware versions) and that age-awareness improved all models to which it was applied[7].

Possible future work is planned to study the scalability of these models, the parameter choice (in order to optimize the models) and on the theoretical properties of the models (proofs of convergence, performance and others). The authors also intend to have users compare the suggested models.

[5] The authors retrieved the 20 highest-ranked articles for both models, and, for both models, all articles have a null *ATC*.

[6] $1 - C$, where C is the controversy. Article controversy is weighted by editor controversy and vice-versa.

[7] Although there is still no non-zero *ATC* article in the first 20 articles ranked by age-aware basic model.

2.2.2 A content-driven reputation system for the wikipedia

Adler and Alfaro[4] developed a reputation system that ranks editors according to how long their editions survive in the articles they edit.

Their content-driven reputation system accounts for text survival and edit (organization and structure) survival, updating the user ratings based on these factors:

- **Text survival**, which measures how much of the *inserted text* is still in the article, updates reputation based on the fraction of inserted text that survived, weighted by the length of the original edition, the (logarithmic) reputation of the editor who kept that fraction of text and by a parameter that needs to be experimentally determined.
- **Edit survival**, which measures how much of the *reorganization* was preserved by future edits, updates reputation based on edit longevity[8], weighted by the edit distance to the previous edit, editor reputation and a parameter that needs to be experimentally determined. If negative, it is also weighted by a "punishment" factor.

This approach was tested in the French and Italian Wikipedias, having performed slightly better than using edit count as reputation[9].

Authors point issues with the system, such as the lack of a knowledge of the Wikipedia conventions and markup language, the unability to track edits across articles. Planned future work includes making the text analysis more precise and studying the performance differences between French and Italian Wikipedias.

They also point out that the used approach is lightweight and efficient, and believe their metrics, unlike edit count, would encourage constructive behavior.

2.2.3 Automatic Vandalism Detection in Wikipedia

Potthast, Stein and Gerling[25] proposed an approach for vandalism detection, defining vandalism detection as a *classification task*, where a set of vandalism indicating features is applied to a given edition. The values of these features are then used to assess a binary classification ("vandalism" or "not vandalism"). This classifier was built with a machine learning approach, being trained with an example corpus.

The authors purpose a set of vandalism indicating features, which rely on language properties and on frequent characteristics of vandalism edits:

- At character level, the deviation from the expected character distribution, the longest repetition of the same character, the compressibility of text (which detects long sequences of the same character[22]) and the upper case to all characters ratio.

[8] Which measures the degree to which the edit was preserved or undone.

[9] Using the current data — it is predicted that if edit count was adopted as an official metric, people would find ways to circumvent the metric, such as doing meaningless, unneeded editions just to boost their reputation. This would make edit count perform worse.

- At word level, the frequency of words in general, and of two word categories: pronouns and vulgarisms, how the edit impacts the number of pronouns and vulgarisms in the article, and the longest word (although in some articles long words are expected, this still catches some nonsense[22])
- Other features are also considered: changes in length, "replacement similarity"[10], "context relation"[11], author anonymity, length of the edit comment and edit count of the author[12].

The authors then evaluated the combination of these features, and it outperformed two existing Wikipedia vandalism detection methods, AntiVandalBot and ClueBot.

This work was extended by Mola-Velasco[22], who provided a modified set of features, and split these into two sets: language-dependent features and language-independent features.

Mola-Velasco does not consider edit count or replacement similarity, splits the upper case ratio in upper-to-lower ratio and upper-to-all ratio, considers the LZW algorithm for the compressibility of text feature and introduces four new features: digit to all characters ratio, non-alphanumeric to all characters ratio, character diversity (compared to the length of the text) and text size increment.

All these new features are considered language-independent, along with the ones inherited from Potthast, Stein and Gerling[25], except the pronoun and vulgarism impact and frequency, which are considered language-dependent.

Similar language-dependent features (impact, frequency) are also added for other word categories: highly biased words, non-vulgar sex words, bad words (including typos and colloquial contractions) and good words, which are not frequently used by vandals (including wiki-syntax elements). A meta-category "All" is also introduced, grouping all categories except "good words".

The author tested different classifiers (C4.5 decision tree, LogitBoost, and Random Forest), using the Weka framework, choosing Random Forest for its performance and stability[13].

Mola-Velasco suggested areas where his work could be enriched: word weighting, considering the *de facto* standard QWERTY keyboard layout when detecting random keyboard hits, analyzing existing robots to extract useful heuristics and experimenting with n-gram language models.

2.2.4 Finding Social Roles in Wikipedia

Welser et al.[34] did some investigation regarding social roles in Wikipedia, proposing four roles and edit history-based metrics to recognize these roles. They also analyzed the impact of user base growth in Wikipedia, to find out if key roles are growing as needed to accommodate the bigger user base.

[10] Which measures, when text is replaced, the similarity of the new and the old texts.
[11] The similarity of edit keywords with article keywords, measures if the edit fits in the context.
[12] That is, how many edits were done by the same user.
[13] Although better values were observed for LogitBoost, these *decreased* when the number of iterations was increased, suggesting instability, while, except for small differences, Random Forest performance increased with more iterations.

The four proposed roles are:

- **substantive experts** editors with extensive knowledge in some topic, possibly even having real world credentials. They contribute substantive content, contributing and resolving conflicts within their areas of expertise. These editors frequently discuss details and check facts.
- **technical editors** editors who carry on some incremental improvements and some maintenance, such as fixing small errors, helping with categorization and building templates.
- **counter vandalism** editors who find vandalized articles, fixing these and sanctioning vandals. Unlike substantive experts, they will have more edits in articles with different, unrelated topics.
- **social networkers** editors who focus on interacting, building ties with other users outside article pages, extensively using user talk pages.

To study the attributes associated with these roles, the authors compiled three sets of contributors:

- **directed** sample of 40 hand-picked contributors, used to document patterns and exceptions. This sample is studied to devise the heuristics.
- **dedicated editor** sample of 1954 contributors whose first edit was made January 2004 or before and who made at least one edit during January 2005.
- **cohort** sample of 5839 contributors who created accounts on January 2005 and made at least one edit during that month. This sample is used to test if adoption of roles changed with time.

The dedicated sample was found to include 33% substantive experts, 10% technical editors, 6% vandal fighters and less than 1% social networkers, while the remaining 50% could not be associated with any role. The cohort sample was found to show a similar relative distribution for these roles.

The authors then conclude that these roles are being replenished, supporting the growth of Wikipedia[14].

After analysis based on the directed sample to assess the structural signatures, some heuristics were proposed that enable the inference of a role from the distribution of the user edits. The authors point out that these heuristics have issues: these do not detect people who play multiple roles and they only cover a non-exhaustive subset of roles.

Possible future work involves refining and testing the heuristics; applying context-based detection (edits on related pages were probably made by substantive experts, while edits on completely unrelated pages were probably made by vandal fighters); crossing roles (if a contributor undoes an edition made by a vandal, then that contributor is possibly a vandal fighter); and monitoring role changes (such as a social networker who decides to start fighting vandalism instead).

2.2.5 Identifying Document Topics Using the Wikipedia Category Network

Schönhofen[28] designed an algorithm to associate Wikipedia categories to a Wikipedia article, based on its title and in the set of all categories in Wikipedia.

[14]The reason behind this question, which is out of the scope of this survey, is that the authors expressed concerns that the survival of systems like Wikipedia depends on people who play key roles.

The algorithm starts by simplifying the original Wikipedia content, reducing it to just articles and redirections, simplifying titles (and merging these as needed), removing administration and maintenance-related categories, removing really small (< 5 articles) and really large (> 5000 articles) categories, and merging "stub categories" with their regular counterparts[15].

Then, articles are simplified, like titles were, removing stop words and performing stemming. Words which do not occur in titles are also removed.

Finally, the author weighted the data using several criteria:

- Words in a document,
- Article titles, which are weighted by the weight of its words, their uniqueness, the number of articles with the title and the percentage of title words that are in the content),
- Articles, which get the maximum weight from associated titles, and
- Categories, whose weights are the sum of weights of their articles[16], weighted by the importance of supporting words for the category relative to its vocabulary. These weights are then re-weighted, in descending order of their weight, in order to make already used words weight gradually less.

Then the n categories with the highest weights can be collected.

The author then carries some experiments: in a specific case, the article named "Analysis of variance", 8 out of the 10 highest ranked categories were, in fact, related to the article.

Generically, using Information Retrieval notation, for the top ranked category ($n = 1$), precision is slightly below 50% with recall between 45 and 50%. Precision and recall were calculated for the n highest ranked categories, with $n \in \{1, \ldots, 10\}$: precision decreases as n increases, reaching a value between 20% and 25% for $n = 10$, and recall only decreases between $n = 1$ and $n = 2$, then increasing with n, reaching 60% for $n = 10$.

Ideas for future work include using the article text and links, and using the hierarchy of categories.

2.2.6 A Breakdown of Quality Flaws in Wikipedia

Anderka and Stein[5] studied the distribution of quality flaws across the English Wikipedia, by compiling a set of tags used to mark flaws in Wikipedia, and dividing the obtained set in twelve flaw categories.

The authors first assessed the existing tags by analyzing the "Cleanup templates" Wikipedia category and the Wikipedia project page "Template messages/Cleanup", which is used by several cleanup tags. After resolving redirects, ignoring some subtemplates (including those meant for documentation, experimentation and inheritance[17]), there were 388 tags.

[15] As exemplified[28], "Economics stub" gets merged with "Economics"
[16] The articles considered here are the ones which have been associated to the category in Wikipedia itself.
[17] Wikipedia pages, especially templates, can be used to build other pages, as "building blocks".

The set was manually inspected to ensure the tags actually refer to cleanup issues (which was found to be the case), with the authors pointing out that, even if it is not possible to prove the completeness of this approach, the set almost surely comprises the most common quality flaws.

Some statistical analysis was carried, showing that tags are mostly used in the encyclopedic content of Wikipedia, where the most tagged topics are those of computer- and belief-related articles (along with a remark that these results reflect the distribution of the flaw tagging effort, not the distribution of flaws).

Flaws were then grouped in twelve categories: Verifiability, Wiki tech, General cleanup, Expand, Unwanted content, Style of writing, Neutrality, Merge, Cleanup of specific subjects, Structure, Time-sensitive and Miscellaneous. This allowed the authors to assess the distribution and type of flaws across Wikipedia, and provides an important contribution towards the automatic detection of those flaws.

2.2.7 Summary

The surveyed works use several metrics, which can be grouped and described as follows:

- **Edit count** Used by four works (Vuong et al.[33], Wikiswarm[46], WikiDashboard[29] and iChase[26]), refers to the use of the number of changes (possibly along time) as a heuristic;
- **Editor anonymity** Only employed by Potthast et al.[25] and Mola-Velasco[22], regards the identification of an author as anonymous, and influencing further decisions based on her anonymity.
- **Number of editors** Concerns using the number of different editors that changed the article as a metric, used by Vuong et al.[33] and Wikiswarm[46].

  Its usage in just one of the four surveyed analysis works is due to the different nature of the works, which focus on incompatible approaches (either because it has no direct meaningful application (such as in Adler et al.[4]), or because it does not apply (as is the case with both Potthast et al.[25] and Mola-Velasco[22], which operate at the revision level)).

  On the other hand, in visualization, its usage in only one of the surveyed works may be the reflection of all other authors' choice on what to visualize. It should be noted that some works may convey a metric to the user as information directly, not using it as part of the process of building their visualization.
- **Number of reverts** The number of reverts that happened on an article. Used by only one work, Revert Graph[30], which is a visualization of the revert network and, as such, relies heavily on revert statistics to build its result.
- **Impact of deletions** The extent to which an edit changes the existing content, only used by Vuong et al.[33].
- **Mutual controversy** Used by Vuong et al.[33], is, as it name states, an assessment of the overall controversy expected to arise from two different users.
- **Compressibility** Involves the application of some compression algorithm and the computation of the resulting *compression ratio*, that is, the ratio between the length of the compressed content and the length of the original, uncompressed content. This value (or its comparison

with a reference value) can then be used to assess if the content deviates from the usual rate found in legitimate texts. Employed by Mola-Velasco[22].

- Other **text statistics**, regarding word and letter frequency are also computed by some works (Potthast et al.[25] and Mola-Velasco[22]), and can be used in the same way as compressibility.
- **Link density** Is also a text statistic, but specific to hyperlinked text, concerning, in this case, the interpretation of Wikipedia markup to find the links present in the content. Other than quality and vandalism, it also presents a measure of how tightly is the article connected to the link structure of Wikipedia. Used by WikipediaViz[11].
- **Similarity** Concerns a more local statistical analysis, comparing the properties of the introduced changes with the scope where they were made, instead of comparing to a more global reference value. Used by Potthast et al.[25], History flow[32, 31] and ThemeRiver[19].
- **Category similarity** Used by Holloway et al.[20], WikiViz[18], ClusterBall[17] and Omnipedia[6] to rule out or relate articles based on their categories, with articles that share categories being considered related.
- **Length** and **age** Stand for those properties of an article and its content. Age is used by Vuong et al.[33], Adler at al.[4], Wikiswarm[46] and Brandes et al.[9]. Length is used by Potthast et al.[25], Mola-Velasco[22], History flow[32, 31] and ThemeRiver[19].

The usage of these metrics across the surveyed works is summarized in table 2.2. It is possible to conclude that unrelated works used different sets of heuristics, except for the *edit count*, *age*, *length*, *similarity* and *category similarity* metrics, which are used by three or more works. It should also be noted that Mola-Velasco[22] shares several heuristics with Potthast et al.[25] because the former work is based on the latter.

This also suggests that, except for these two works [22, 25], the different approaches are not redundant. We also found no incompatibility between any metrics.

A couple other indirectly related metrics are not discussed above nor listed in table 2.2: the heuristic proposed by Welser et al.[34] for social roles relies on the distribution of edits through several kinds of pages, and the topic identification algorithm[28] uses the article title and the set of Wikipedia categories. Likewise, Anderka and Stein[5] rely on the set of categories, along with the set of templates, to identify quality flaws.

These metrics are applied at different levels:

- Some metrics refer to **sets of revisions**, instead of a specific revision. This includes edit count, number of editors, number of reverts and article age;
- Category similarity operates on a **contribution-specific level**, that is, on the part of the revision content that was changed by that revision;
- Mutual controversy, unlike the other metrics, focuses on **users**, not on articles;
- The remaining metrics (impact of deletions, length, compressibility, text statistics and author anonymity) are appliable to a **single revision** or a finer granularity level (for example, at the section or contribution levels).

	Analysis				Visualization											
	Vuong et. al[33]	Adler et al.[4]	Potthast et al.[25]	Mola-Velasco[22]	Wikiswarm[46]	History flow[32, 31]	ThemeRiver[19]	Revert Graph[30]	Brandes et al.[9]	Holloway et al.[20]	WikiDashboard[29]	WikipediaViz[11]	iChase[26]	WikiViz[18]	ClusterBall[17]	Omnipedia[6]
Edit count	✓		✓		✓						✓		✓			
Editor anonymity			✓	✓												
Number of editors	✓				✓											
Number of "reverts"								✓								
Impact of deletions	✓															
Mutual controversy	✓															
Age-based	✓	✓			✓				✓							
Length			✓	✓		✓						✓				
Compressibility				✓												
Text statistics			✓	✓												
Similarity			✓			✓	✓									
Link density												✓				
Category similarity										✓				✓	✓	✓

Table 2.2: Comparison of metrics and their use in several works.

The suitability of these categories of metrics for a specific visualization is dictated by the choices made regarding the type of visualization, and the level at which it will operate. For example, it may be difficult to convey metrics on sets of revisions in a way other than textual information, unless the visualization is conceived in a way that allows such information to be presented using an adequate approach (such as a time-series visualization).

Metric suitability is also restricted by the computational costs of the involved computations, by which metrics can be split into groups:

- metrics based on networks of articles (category similarity)
- metrics based on networks of users (mutual controversy)
- metrics based on revision content (impact of deletions, length, compressibility and text statistics)
- metrics based on revision metadata (author anonymity)
- metrics based on revision metadata, at an article level (article age, number of reverts, editions and editors)

Of those, only the network-based metrics pose a greater challenge, as all the others operate at the article level, and can thus be easily computed given a set of revisions.

Chapter 3

Metrics Assessment

In chapter 2, previous works were separated in two groups, one for visualizations, other for metrics. We decided to apply this distinction to our own work, dividing the overall project in two parts: a visualization, concerned with the presentation of already computed data to the user, discussed in chapter 4 and a metrics computation component, which we present in the current chapter. This defines the high-level architecture of the system, presented in figure 3.1.

Figure 3.1: High-level representation of the proposed system.

The metrics computation component concerns the delivery of processed data requested by the visualization, after the execution of computation and processing steps on original data retrieved from a MediaWiki instance.

It is, thus, in the design of this component that decisions are made on *what* to show in the visualization. The design of the visualization component will be concerned with *how* to show the information made available by the component we describe in the current chapter.

In section 3.1, we start by choosing and defining a set of metrics to depict in the visualization, then detailing how and where is the required information obtained from.

Then, in section 3.2, we describe the adaptation of the revision content to a format which can be used directly by the visualization.

Finally, we present the overall architecture of the component, along with the design decisions driving some of the choices, in section 3.3.

3.1 Computing Statistics

In order to chose the metrics to be employed by this component, we first analyzed the visualization works summarized in section 2.1.13. Considering our goal of visualizing information from the log of changes, or, more precisely, the evolution of articles through time, and the considerable amount of information involved, we chose to focus our visualization at the article level.

We believe this choice will allow us to decrease the amount of information what needs to be depicted, allowing the visualization to convey more details about the same article. It should be noted that, not only is the entire English Wikipedia database very large (encompassing more than 5 TiB of data), but also that single articles are still too large to be directly analyzed (for example, the English Wikipedia article on "Elephant", at the time of its 5113rd revision, involved 300 MiB of data.

With this choice in mind, and considering the set of metrics collected in section 2.2.7, we devised a set of statistics:

With the concern to keep the system interactive, due to the data volume concerns mentioned above, and considering the vast size of the network of articles and categories, we chose to avoid inter-article or user-based metrics that cannot be computed from the article data (thus excluding mutual controversy and category similarity).

With the same real-time concern in mind, we decided to use statistics that could be computed out of the content of a specific revision (that is, independently of the content of past and future revisions) and, eventually, from a fixed reference value or set, to be used for all articles. This led to the inclusion of the impact of deletions, length, compressibility, text statistics and link density, while excluding similarity.

Finally, the identification of reverts is directly obtainable from the revision comment, with its number, along with the number of editors, the edit count and the article age being implied in the visualization time-series approach, and author identity being made available as plain information, enabling the user to assess the author anonymity[1].

These decisions led to the chosen metrics, a subset of the list shown in table 2.2. These are edit count, author anonymity, number of editors, number of reverts, impact of deletions, age (revision date), length, compressibility, text statistics and link density. These metrics, summarized in table 3.1, along with the revision URL, define the set of metrics from which the visualization data will be computed.

Metrics	**How to obtain**
Edit count	Implied in visualization (time series)
Author anonymity	Revision properties
Number of editors	Implied in visualization (time series)
Number of reverts	Look for "rv" in edition comments
Impact of deletions	Compute the ratio between the number of words deleted and the number of words contributed to the article, and compare with a reference value.
Age	Implied in visualization (time series)
Length	Text length.
Compressibility	Apply a set of compression algorithms and get the average compression rate and compare with a reference value.
Text statistics	Compute a set of character- and word-based ratios and frequencies over the article content and compare with a reference value.
Link density	Count the number of linked words and compare with a reference value.

Table 3.1: List of metrics used by the proposed system.

[1] Authors are identified either by their username, if they are registered in Wikipedia, or by their Internet Protocol address, if they are not registered, i.e., if they are anonymous users.

The processed data, to be served to the visualization, encompasses six directly assessable article properties:

- **date** and **revid** (the "revision id");
- The **URL** of the revision in the web interface to the MediaWiki instance under analysis;
- Text **length**, the length of the revision content;
- Whether the revision is a **revert**;
- The revision **author**.

3.1.1 Other Values

The visualization is also provided with **quality**, **vandalism**, **controversy**, **change impact** and an indication of whether the edit **is a deletion**. These metrics are not directly obtainable from the MediaWiki software, and must be computed using intermediate values, which we will also introduce.

These metrics are presented below, and exemplified with their values from the English Wikipedia revisions 30603257[2], its successor 30603285[3] (an act of vandalism, where several repetitions of the expression "ELEPHANTS R COOLOOOOOOOOOOOHHHHH[...]HHH" were added to the article body) and 30603339[4] (which reverted the vandalism). All these revisions are from the article on "Elephant", and are referred, respectively, as "revision 1", "revision 2" and "revision 3" below.

We start by defining two auxiliary metrics:

- The *number of changes*, c, computed from the set of word-level differences between the current revision and the previous revision. Given the number of deleted words d and the number of added words a, it is given by the sum of both:

$$c = d + a \tag{3.1}$$

- The *deletion impact*, i_d which measures the ratio between the number of deleted words and the number of changes[33]. With c and d as defined in equation (3.1), it is given by:

$$i_d = \frac{d}{c} \tag{3.2}$$

The **change impact**, i_c, is then defined, in a way similar to the deletion impact[33], as the ratio between the number of changes c and the sum of the number of words of the previous revision (n_p) and of the current revision (n_c):

$$i_c = \frac{c}{n_p + n_c} \tag{3.3}$$

[2] `http://en.wikipedia.org/w/index.php?oldid=30603257`
[3] `http://en.wikipedia.org/w/index.php?oldid=30603285`
[4] `http://en.wikipedia.org/w/index.php?oldid=30603339`

All ratios above are defined so that their value is zero when their denominator is null. When there is no previous revision, values are set to reasonable, expected defaults:

- As there are no deletions and, thus, no changes (as nothing was changed), we define both the number of changes c and the deletion impact i_d to be null, i.e.,

$$c = 0 \tag{3.4}$$

$$i_d = 0 \tag{3.5}$$

- On the other hand, as this edition introduced a new article or section, the change impact i_c is set to its maximum, 1:

$$i_c = 1 \tag{3.6}$$

The information on whether the edition **is a deletion** is then set to be true if and only if the deletion impact is the highest possible, that is, when $i_d = 1$.

$$\text{deletion?} = \begin{cases} \text{yes} & \text{iff. } i_d = 1 \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

The values of the metrics defined above for the three revisions of the "Elephant" article are presented in table 3.2. It is worth noting that the deletion impact allows us to correctly identify revision 3 as a deletion of content, which is expected, as it is a revert, where the author just removed the text added in revision 2. Revision 1 was also a revert, as noted in its revision comment ("rv vandalism").

revision	1	2	3
d	—	0	17710
a	—	17710	0
c	0	17710	17710
i_d	0	0	1
i_c	1	≈ 0.5209	≈ 0.5209
deletion?	no	no	yes
revert?	yes	no	yes

Table 3.2: Values of the auxiliary metrics for the three chosen revisions of "Elephant".

For the remaining heuristics, **quality**, **vandalism** and **controversy**, some more intermediate values are computed, which are then compared with an average value: the average of these auxiliary metrics for one thousand revisions from the "recent changes list" as of June 13, 2012 12h12 UTC, which we assessed ourselves.

Compressibility: The compressibility of texts is the compression ratio obtained after applying a compression algorithm[22], in our case the DEFLATE algorithm. For the chosen corpus, this is $r_{c_0} \approx 0.362$.

Let l be the uncompressed article length and $l_{DEFLATE}$ its length after the application of the DEFLATE compression algorithm. Then, the compressibility ratio, r_c, is defined as:

$$r_c = \frac{l_{DEFLATE}}{l} \tag{3.8}$$

And the compressibility score s_c, measuring the deviation from the reference ratio, is defined so that its value falls in the range $[0; 1]$,

$$s_c = \begin{cases} 1 - \frac{|r_c - r_{c_0}|}{r_{c_0}} & \text{if } r_c < 2 \cdot r_{c_0} \\ 0 & \text{otherwise} \end{cases} \tag{3.9}$$

Character frequency: Measures what are the most frequent characters of the text[25, 22]. The nine most frequent characters were assessed from the reference corpus[5] $chr_0 = (a, e, h, i, n, o, r, s, t)$. This set is compared to the set of the ten most frequent characters of each text (chr), and the length of the intersection of both sets is used to compute the character frequency score, $s_{chr} \in [0; 1]$, which measures how similar both sets are:

$$s_{chr} = \frac{length(chr \cap chr_0)}{9} \tag{3.10}$$

Word frequency: frequencies of several categories of words were considered[25, 22], with their scores being computed the same way character frequency is:

- **"good words"**, w_g, are the most frequent words in the chosen corpus: 'and', 'is', 'it', 'an', 'as', 'at', 'in', 'from', 'for', 'this', 'to', 'which', 'has', 'was', 'be', 'his', 'that', 'with', 'by', 'he', 'a', 'on', 'of', 'were', 'the', 'first', 'or' and 'are';
- **"biased words"**, w_b, were extracted from the Wikipedia Manual of Style[41]: 'legendary', 'great', 'eminent', 'visionary', 'outstanding', 'leading', 'celebrated', 'cutting-edge', 'extraordinary', 'brilliant', 'famous', 'renowned', 'remarkable', 'prestigious', 'world-class', 'respected', 'notable', 'virtuoso', 'cult', 'racist', 'perverted', 'sect', 'fundamentalist', 'heretic', 'extremist', 'denialist', 'terrorist', 'controversial', 'supposed', 'purported', 'alleged', 'accused', 'so-called', 'notably', 'interestingly', 'clearly', 'certainly', 'fortunately', 'happily', 'unfortunately', 'tragically', 'untimely', 'reveal', 'expose', 'explain', 'find', 'note', 'observe', 'insist', 'speculate', 'surmise', 'claim', 'assert', 'admit', 'confess' and 'deny';
- **"offensive words"**, w_o, were also taken from Wikipedia[39]: 'arse', 'asshole', 'asswipe', 'bitch', 'bollocks', 'breeder', 'bugger', 'bullshit', 'cunt', 'damnation', 'faggot', 'feck', 'frak', 'fubar', 'fuck', 'git', 'motherfucker', 'nigger', 'pissant', 'pussy', 'shit', 'slut', 'spastic', 'twat' and 'wanker'.

The set of the fifteen most common words, w, is extracted and, in the same way as in equation (3.10), is intersected with each of the three sets above, with the length of the intersection being used to assess the good, biased and offensive word scores (respectively, s_{w_g}, s_{w_b} and s_{w_o}, all of which are within the range $[0; 1]$):

[5] Although the specific frequency will be language-dependent, the reference set will still work for other languages which use the Latin alphabet. Unlike the word scores we will introduce later, the set of frequent characters will be similar across languages, and we do not consider the position of characters inside the set.

$$s_{w_g} = \frac{length(w \cap w_g)}{15} \tag{3.11}$$

$$s_{w_b} = \frac{length(w \cap w_b)}{15} \tag{3.12}$$

$$s_{w_o} = \frac{length(w \cap w_o)}{15} \tag{3.13}$$

Link ratio: Is defined as the ratio between the number of links n_l and the number of words n_w[11],

$$r_l = \frac{n_l}{n_w} \tag{3.14}$$

where the number of links is obtained by looking for MediaWiki link markup in the revision content.

The reference link ratio, r_{l_0} was also computed from the corpus, whose articles contain a total of 147115 links and 2802251 words, and thus a link ratio of $r_{l_0} \approx 0.052$.

The link score s_l can then be defined as:

$$s_l = \begin{cases} 1 - \frac{|r_l - r_{l_0}|}{r_{l_0}} & \text{if } r_l < 2 \cdot r_{l_0} \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

Composite metrics: The three remaining metrics are computed using several of the values defined above and according to some expectations, explained below, based on examples from the three revisions chosen above, whose values for the remaining metrics above are shown in table 3.3. As revision 3 reverted the content of the article to that of revision 1, and all these metrics are content-based, they show, as expected, the same values for both revision 1 and revision 3.

The formulas for these metrics are based on our expectations for articles showing good quality, for controversial issues and for the content introduced by acts of vandalism.

This follows an approach similar to that of Mola-Velasco[22], where several metrics are aggregated together using techniques from machine learning, although we opted by using an unweighted average of the values, values, instead of gathering weights using a supervised learning step, which would require training with a specific corpus, possibly tying the resulting weights to a specific family of articles or set of languages, in a way that hampers the application of its results to article families or languages which were not covered by the chosen corpus.

revision	1	2	3
r_c	≈ 0.3951	≈ 0.0505	≈ 0.3951
s_c	≈ 0.8647	≈ 0.1452	≈ 0.8647
s_{chr}	1.0	≈ 0.8889	1.0
s_{w_g}	≈ 0.7333	≈ 0.7333	≈ 0.7333
s_{w_b}	0.0	0.0	0.0
s_{w_o}	0.0	0.0	0.0
r_l	≈ 0.0254	≈ 0.0080	≈ 0.0254
s_l	≈ 0.4888	≈ 0.1540	≈ 0.4888

Table 3.3: Values of the remaining auxiliary metrics for the three chosen revisions of "Elephant".

- **Quality (q)**: We expect normal, encyclopedic, on-topic content to show similar features and heuristics, regarding the distribution of characters, the used words and the number of hyperlinks.

  Thus, we define a *quality* metric, q, with values ranging from 0 to 1, averaging a set of scores, defined above, which measure how close is the current text to the average normal text, regarding the most frequent words found in normal text ("good words"), the character distribution, the link ratio and the compressibility score. We also measure the occurrence of offensive or biased words, penalizing texts with such occurrences.

  This *quality score* is defined as follows:

$$q = \frac{s_{w_g} + (1 - s_{w_b}) + (1 - s_{w_o}) + s_c + s_l + s_{chr}}{6} \tag{3.16}$$

  As the vandalism in revision 2 involved the repetition of a specific short sentence, it did not affect the word scores, even if the introduced sentence deviates significantly from the expected natural language character pattern. Although the character repetitions are not enough to affect the character frequency, the repeated sentence allows the compression algorithm to achieve a significantly better compression ratio, thus significantly reducing the compressibility score. A similar drop occurs in the link score due to the introduction of the big block of text, as it has no links.

  This leads to the desired quality value changes: revision 1 and 3 share a quality measurement of $q_1 = q_3 \approx 0.8551$, while revision 2 is assessed as having $q_2 \approx 0.6526$.

- **Vandalism (v)**: When users vandalize a page, they are not interested in contributing with encyclopedic and on-topic content. They are, in fact, interested in the opposite: contributing off-topic non-encyclopedic content, with the sole purpose of disturbing Wikipedia and the work of its contributors.

  Not only is the content of vandal contributions unrelated to the article, those are also frequently made of meaningless text, not following the usual distribution of words in the language, or even not featuring proper words at all. Parts of vandalism contributions will feature repetitions of the same character or sequences of random characters.

  This is expected to negatively affect the good words, compressibility and character frequency scores, thus showing that the contributed text *deviates* from what would be considered quality text. Vandals may also choose to disturb the encyclopedia by adding a great amount of links to unrelated articles, negatively affecting the link score.

  Some acts of vandalism may involve not the introduction of meaningless, random sequences of characters, but the introduction of offensive words, which translates into an increase of the "offensive words" score.

  Once again, these scores are averaged to devise the metric:

$$v = \frac{s_{w_o} + \left(1 - s_{w_g}\right) + (1 - s_c) + (1 - s_l) + (1 - s_{chr})}{5} \tag{3.17}$$

  Although, as mentioned above, the word scores were not affected by this act of vandalism, the fluctuations of the remaining scores, including character frequency, do indeed match our expectations regarding the deviation from the normal, expected structure and statistics of natural language, resulting in slightly more than twice the vandalism score for the second revision, and in low values for the first and third revisions:

$$v_1 \approx 0.1738 \tag{3.18}$$

$$v_2 \approx 0.4168 \tag{3.19}$$

$$v_3 \approx 0.1738 \tag{3.20}$$

- **Controversy** (c): On the other hand, when users are interested in explicitly supporting a specific point-of-view, possibly in violation of Wikipedia's "Neutral point of view" policy, they will write text that shows properties similar to those of good text, showing a similar distribution of words and characters. But, unlike good text, an attempt to support a specific point of view may involve the usage of known-bad words: words that introduce a bias and that are frequently used to build unsupported claims: "biased words".

  Unlike vandals, controversial editors will make an attempt to build content that looks like good, regular content, distancing themselves from vandalism, thus we expect the deletion impact to remain low.

  Therefore, we define *controversy*, c, as the average:

$$c = \frac{s_{w_g} + s_{w_b} + s_{chr} + (1 - i_d)}{4} \tag{3.21}$$

  As the chosen example is an act of vandalism, we did not expect meaningful changes in controversy, other than a decrease in the third revision, as it is a deletion, and has a high deletion impact:

$$c_1 \approx 0.6833 \tag{3.22}$$

$$c_2 \approx 0.6556 \tag{3.23}$$

$$c_3 \approx 0.4333 \tag{3.24}$$

Special cases: When observing the results of the formulas above, vandalism (3.17) and controversy (3.21) would present abnormally large values for small sections. This was found to be due to the scarce content of such small sections.

For example, the last section of revision 477831327[6] from the English Wikipedia article on "Steel Bank Common Lisp", which contains the section header, three external links and MediaWiki category tags, lacks enough words to achieve a significative good words score ($s_{w_g} \approx 0.0667$). Its scarce content also affects the compressibility ratio leading to a slightly lower score, $s_c \approx 0.6292$. The most affected metric is the link score, as this section only contains links, thus reaching the lowest score possible, $s_l = 0$.

Due to these effects, the assessed composite metrics present higher values than desired:

[6] `http://en.wikipedia.org/w/index.php?oldid=477831327`

$$q \approx 0.5789 \tag{3.25}$$

$$v \approx 0.5053 \tag{3.26}$$

$$c \approx 0.4611 \tag{3.27}$$

In order to avoid this undesired effect of the lack of content, these two metrics were redefined when the length l of the content is under 1000 characters, a length we experimentally determined to be a good boundary between regular sections and the smaller sections which do not have enough content.

In these cases, the metrics are weighted by a factor $f \in [0.2; 1]$, proportional to the article length (empty articles have $f = 0.2$, while 1000 characters-wide articles will have $f = 1$),

$$f = 0.2 + 0.8 \cdot \frac{l}{1000} \tag{3.28}$$

With this weighting in place, the two composite metrics present less aggressive values,

$$v' \approx 0.3161 \tag{3.29}$$

$$c' \approx 0.2885 \tag{3.30}$$

The three composite metrics were also redefined in the case the MediaWiki being accessed was other than the English Wikipedia, as several of the scores used in (3.17), (3.21) and (3.16) depend on specific features of the English language[7].

First, equivalent word categories are defined for the French, Portuguese, and Spanish Wikipedias, by assessing the most frequent "good words" in a similar way, and using word lists from the Wikipedia project itself or from Wiktionary. The complete lists are presented in appendix B, from which we show some examples:

- French Wikipedia
  - **"good words"**: 'une', 'qui', 'que', 'se', 'ce', 'aux', …;
  - **"offensive words"**: 'beauf', 'con', 'gouine', 'métèque', 'nègre', 'social-traître', …;
  - **"biased words"**: 'souvent', 'généralement', 'inégalé', 'prétendre', 'essentiellement', 'théorie', 'dictateur', 'auteur', …;
- Portuguese Wikipedia
  - **"good words"**: 'de', 'em', 'que', 'para', 'por', 'dos', 'as', 'mais', 'ser', …;
  - **"offensive words"**: 'babaca', 'bardamerda', 'barrote', 'bedamerda', 'berdamerda', 'bunda', 'burro', 'cabaço', 'cacete', 'cachorra', 'cadela', 'mamado', 'maricas', 'merda', …;

[7] These are, in fact, similar to the features Mola-Velasco[22] classified as "language dependent" (as discussed in section 2.2.3).

  - **"biased words"**: 'lendário', 'grande', 'eminente', 'visionário', 'notável', 'líder', 'célebre', 'extraordinário', 'terrorista', 'claramente', 'certamente', 'afortunadamente', 'infelizmente', 'tragicamente', 'negou', …;

- Spanish Wikipedia
  - **"good words"**: 'en', 'el', 'su', 'una', 'al', 'es', 'sus', 'entre', 'este', 'esta', …;
  - **"offensive words"**: 'caraja', 'carajo', 'correrse', 'coño', 'criatura', 'culero', 'culiado', 'culicagado', 'culo', 'maricón', 'marihuanero', 'mazo', 'mear', 'mes', 'mierda', 'mojar', 'mojarse', 'mojón', 'mostacero', 'nabo', 'nardo', 'orto', 'paja', …;
  - **"biased words"**: 'afirmó', 'pretendido', 'aunque', 'naturalmente', 'afortunadamente', 'curiosamente', 'tristemente', 'trágicamente', 'escándalo', 'polémica', 'teóricamente', 'secta', ….

Then, these metrics were redefined for all other MediaWiki sites, for which we do not provide corresponding word lists. Those alternative definitions are as above, except for the word scores, which are not considered:

$$q_i = \frac{s_c + s_l + s_{chr}}{3} \tag{3.31}$$

$$v_i = \frac{(1 - s_c) + (1 - s_l) + (1 - s_{chr})}{3} \tag{3.32}$$

$$c_i = \frac{s_{chr} + (1 - i_d)}{2} \tag{3.33}$$

3.1.2 Retrieving Data From Wikipedia

The values presented above are computed from MediaWiki metadata and content, as follows:

- In order to compute the revision date, we need the **revision timestamp**;
- The **revision id** is a metadata property itself, and enables us to build the revision URL;
- We need the **revision comment** in order to tell whether the revision is a revert (which we assess by looking for the "rv" string in the revision comment);
- The **author's username** is also a metadata property;
- Length, number of changes, deletion impact, change impact and the text statistics (compressibility, character frequency, word frequency and link ratio) are all computed from the **revision content**.

These five values will have to be retrieved from the MediaWiki API, which provides several services, whose request formats and explanations are listed when the API is accessed with no parameters[8].

[8]That is, when one accesses the API endpoint through its address without specifying any parameters, for example, by accessing the endpoint in a web browser. The endpoint for the English Wikipedia is at `http://en.wikipedia.org/w/api.php`.

In our case, we are interested in the revision-related properties listed above. These properties are provided through the API `query` action, which allows the retrieval of several article-related data ("properties"), including revision-level data.

When requesting revision-level data (`prop=revisions`), it is possible to specify which revision-level properties are returned. In our case, we are interested in

- the timestamp of the revision (`timestamp`)
- the revision id (`ids`)
- the revision comment (`comment`)
- the author's username or Internet Protocol address (`user`)
- the revision content (`content`)

This request returns the revisions in order, by default starting with the newest revision, but this order can be explicitly stated through the `rvdir` parameter (the value `newer` returns revisions in increasing id order (i.e. oldest first), and `older` exhibits the default behavior (decreasing id order)).

If we were accessing the API directly and wanted to retrieve the content of the English Wikipedia main page, we would issue the following request:

```
https://en.wikipedia.org/w/api.php?action=query&prop=revisions&titles=Main%20Page
    &rvprop=ids|content|user|comment|timestamp&rvdir=newer
```

For which the server would return the answer as a web page with XML code. There are several formats available, including plain XML, which we could request and parse. But, in order to simplify our system and avoid the need of dealing directly with the API and its HTTP interface, we decided to use the `mwclient` Python package to communicate with the MediaWiki API.

`mwclient` provides the results in a Python-friendly way, opting by Python generators instead of plain lists, thus providing one of the main benefits of Python generators: delayed evaluation. Revision data will not be retrieved at once, `mwclient` will request more data as needed. `mwclient` also makes the revision data available as a Python dictionary, easing its manipulation.

Special cases: Although the API serves revisions by chronological order, there are some special cases where, for some reason, specific revisions have out-of-order timestamps.[9] In those cases, when the timestamp is found to be older than the one from the previous revision, we inherit the previous revision timestamp.

MediaWiki has support to hide data from public logs. In extreme cases, Wikipedia administrators may use this feature to hide problematic data or information they are forced to redact by law (such as in English Wikipedia revisions 347569077[10], from the article on the "Texas Instruments signing key controversy", and 356614223[11], from the article on the Eurovision Song Contest winner "Lena

[9] Examples are English Wikipedia revisions 386093764 (`http://en.wikipedia.org/w/index.php?oldid=386093764`), from the article on the "COBOL" programming language, and 10991081 (`http://en.wikipedia.org/w/index.php?oldid=10991081`), from an archive of the talk page for the article on "Mainland China". Some of these cases are documented at `http://en.wikipedia.org/wiki/User:Graham87/Page_history_observations#Strange_times_reported_in_diffs`, and are possibly a consequence of out-of-sync server clocks.

[10] `http://en.wikipedia.org/w/index.php?oldid=347569077`

[11] `http://en.wikipedia.org/w/index.php?oldid=356614223`

Meyer-Landrut"). In these cases, administrators can hide the revision content, the author's username or both. In the `mwclient` result, this translates into missing entries in the resulting revision structure.

API Etiquette: This component will download a considerable amount of data. Other than choosing a cache-based design, we also try to minimize the impact of this data flow on the Wikipedia servers by following the official API Etiquette[21], according to which

- "Requests should be serialized rather than done in parallel", a property our code was written to satisfy.
- "The `maxlag` parameter should be used to refrain from doing the requested action when the server is under high load", already respected by `mwclient`.

3.2 Content Parsing

Our visualization will also allow users to access the content of each revision. This is done using a cached, low-fidelity version of the content, which we store in a cache.

The revision content, as retrieved from MediaWiki, is written using MediaWiki markup. As the visualization will use the content in a way intended for users to read it, some markup constructs have to be handled before the content is used for display.

In this process, some trade-offs were made: regular expressions are not the best solution to parse MediaWiki syntax, but are several orders of magnitude faster than more complete approaches. We chose producing some usable content in a short time rather than increasing even more the article processing time by trying to get a perfect, faithful rendering of MediaWiki markup.

This decision is better understood if we focus on more popular articles, which receive more attention from contributors, may have many sections and will have thousands of revisions, which will have to be separately processed by this component.

Therefore, we employ the following translation steps, whose result is shown in figure 3.2 for a short example of MediaWiki markup:

1. HTML headings are added. In the MediaWiki markup language, sections are created by surrounding the section title with equal signs, using the same number of signs on both sides.
2. Then list bullets (asterisks at the beginning of a line) are converted to HTML list items (`<LI>`).
3. Images and other generic embed media are translated as `<IMG>` and `<OBJECT>` tags.
4. Wikilinks, links to other articles inside the same MediaWiki instance, are then translated to the HTML anchor element (`<A>`).
5. External links are also translated to HTML anchors, using an arrow symbol (⇗) as the anchor text if there is no link text in the original MediaWiki markup.
6. Text emphasis is then translated to the HTML emphasis tag (`<em>`).

7. MediaWiki templates, which may be too complicated and time-intensive for regular expression handling, are hidden using HTML `<SPAN>` elements with the CSS display style "none".

8. Finally, paragraphs are enclosed in pairs of HTML `<P>` and `</P>` tags.

```
Example article
== Section One ==
''Once'' upon a time...

* There was a man;
* He sailed to sea.

== Section Two ==

[[Eleanor Rigby|Father McKenzie
    ]]

[[Image:Beethoven symphony 5
    opening.svg]]

[http://www.applerecords.com/]
```

Example article

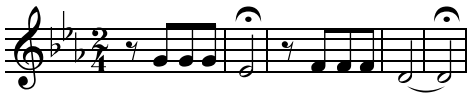
Section One

Once upon a time...

- There was a man;
- He sailed to sea.

Section Two

Father McKenzie



[edit]

Figure 3.2: MediaWiki markup parsing example: source code (left) and the resulting HTML, as rendered by WebKit.

Computation of differences: The resulting HTML is then split by paragraphs, and differences between the current and previous revision are computed from the lists of current and previous paragraphs, using Python's own `difflib`.

HTML diffs are generated, in a way similar to the plain HTML content, but where removed paragraphs are highlighted with a red background and introduced ones with a green background. An example of a change from figure 3.2 is shown in figure 3.3, where the second item in the bulleted list was changed from "He sailed to sea." to "He sailed to the land of submarines.".

Differences are computed at the paragraph level, with an entire paragraph being marked as having been changed if some word inside it changes. Lists, such as the one shown in figure 3.3, are regarded as a single paragraph.

3.3 Request Processing

This component was implemented as a Python package, which provides the values used by the visualization, through a HTTP interface (described in section 3.3.1), and which computes these values based on data provided by a MediaWiki instance (section 3.3.2). These values may be already cached (section 3.3.3).

These subcomponents are orchestrated in order to compose the whole metrics assessment component, as shown in figure 3.4.

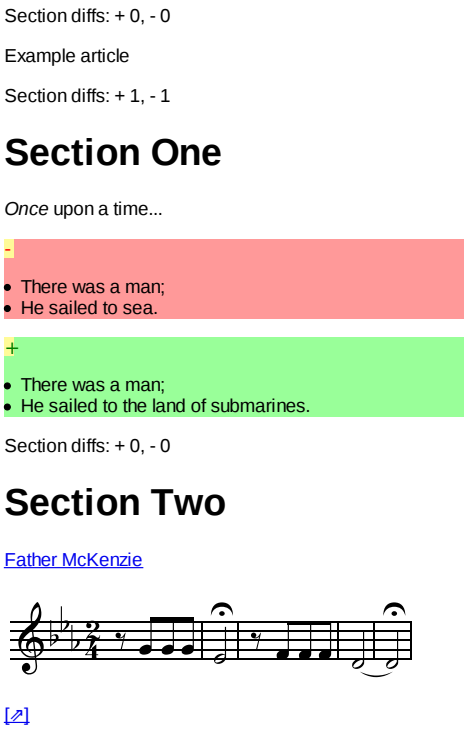


Figure 3.3: MediaWiki differences parsing example, as rendered by WebKit.

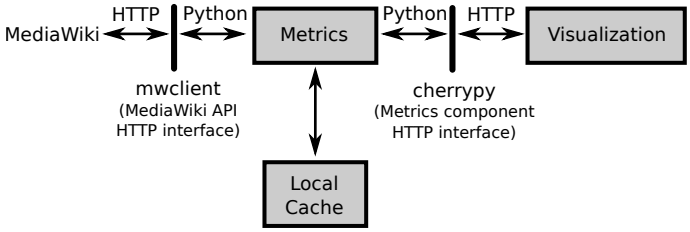


Figure 3.4: High-level representation of the proposed system.

3.3.1 Metrics Component HTTP Interface

The Python package that comprises this component allows the retrieval of content, content changes (diffs) and statistics, either article- or section-wide, for a specified time range, either as a Python object or as JSON-encoded data[12], and offers those features through a HTTP interface (built on top of `cherrypy`[13]), which is used by the visualization to retrieve the needed data.

As far as the visualization component is concerned, this part provides, given an article, its content and content changes, along with metrics (for each revision, either article-wise or by section) and the article talk page activity.

All these requests are made on specified time ranges, and the results are filtered to include only the entries which are in the range.

The web interface provides these data through three services: retrieval of metrics (revision-wide or for all sections) and talk page activity, retrieval of revision content and retrieval of revision content differences.

[12] JavaScript Object Notation, a textual representation for objects in JavaScript.
[13] A Python package designed to ease the deployment of HTTP endpoints for Python functions.

An additional service is also made available to provide the visualization with some feedback on the article processing, by disclosing the progress (the number of already processed revisions along with the total number of revisions).

Except for the retrieval of content and content differences, which are similar in nature, all these requests are for entirely different purposes and are expected to be done at different rates:

- Progress information will only be requested when an article is processed for the first time;
- Metrics and talk page activity are requested each time the visualization has to be redrawn;
- Content and content differences will be requested several times for the same screen, if the visualization is set to show the revision content, as this is updated as the user moves the mouse pointer over the visualization.

The different request rates will be exploited when addressing the design of a cache to account for the time complexity of the required computations (section 3.3.3), although it is ignored when choosing how to process articles (section 3.3.2), as, despite the different rates, the responses to these requests are computed from the same source: revision content and properties.

3.3.2 Article Processing

When an article is requested, its content is requested from the API and the computations described in section 3.1 are carried out on the data returned by the API.

As we want the user to be able to focus on a granularity level deeper than the article level, we also provide section-level metrics, for which the computations are applied over the same revision data, but with a subset of the original content, corresponding to a specific section.

Statistics and diffs are computed based on a specific revision, together with the result of processing the previous revision. That is, computations are mostly independent, a property we will use in section 3.3.3 to optimize the time complexity exhibited by this component.

Section tracking: When generating section-level statistics, we had to account for changes in the section order, adopting an approach which causes sections to be reordered to match the order present in the previous revision, making our code robust against section moves.

This matching is done by comparing the content, split by sections, with the previous content of all sections found so far. First, we pair identical sections together, and then we pair the remaining sections with the closest matches, if these matches are close enough. Finally, all the remaining sections, which have not been matched by the previous steps, are considered to be new sections.

Text closeness is measured using a proximity score provided by Python's `difflib`[16], defined as the ratio between the number of words shared by both texts and the total number of words in the two texts. In order to avoid matching new sections with unrelated deleted sections, if the best score found for a section is lower than 0.3, then the algorithm will refuse to match the section.

Simply splitting by sections without this tracking would lead to undesired scenarios such as in an article presenting the following sequence of events:

1. Revision 1: Ben Bitdiddle creates the article, with sections "Introduction" and "Controversial ideas"
2. Revision 2: Alyssa P. Hacker splits "Introduction" in "Introduction" (the section) and subsections "Early history" and "Today"

In the absence of the tracking mechanism, the section "Controversial ideas" from the first revision (the second section) would be paired with the section "Early history" of the second revision, because it is the second section from the second revision, even if the section "Controversial ideas" was not deleted and may even have not been changed at all (figure 3.5, left), while the content-aware tracking method above would correctly track the "Controversial ideas" section (figure 3.5, right)

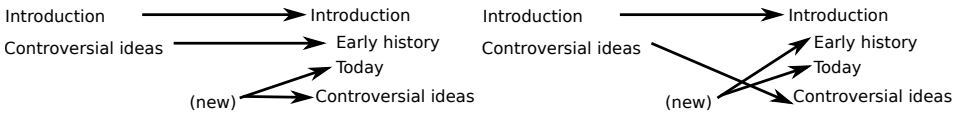


Figure 3.5: Sectioning strategies: naive (left) and content-aware (right).

Granularity of processing: As mentioned in section 3.3.2, revisions are processed in an almost independent way, which allows for different approaches: processing the entire article *at once* or processing each requested range on the first time it is requested, adding the *partial results* to the cache.

Processing ranges on demand would be faster for small subsets of the revision history, but we expected users to explore large parts of the article history, if not its entirety, so that they could get a "big picture" of the article evolution along its existence.

It would also require additional tracking code as, while revisions are identified by their numeric id, requests are made for time ranges. Such code would either retrieve some metadata (revision ids along with their timestamps) for the entire history at once, or would have to identify already cached time ranges, retrieving and processing only the required data.

As a single visualization screen would likely involve many revisions, on-demand computation would also cause an operation as simple as panning the plot to block the visualization for several minutes, were results computed on demand. In this case, concentrating the entire processing steps on a single wait time would seriously improve the user experience — despite a longer initial wait, all further operations would be seamless, as far as this component is concerned.

And, last but not least, already computed data would have to be reevaluated to accommodate changes introduced by the newly computed data.

While, for the diffs and revision statistics, this could be as simple as getting the preceding revision, using its value, and then discarding it, the code also tries to keep track of sections, making the visualization resistant to section moves or subsectioning.

Even if this tracking mechanism were adapted to work with partial, on demand retrieval of statistics, there would be no way to guarantee a consistent section ordering across independent executions of the

system, as the final section ordering would depend on the ordering of sections inside the first requested time range[14] — the only way to ensure a consistent order would be to process the entire article at once.

In order to avoid these issues, we decided to process the entire article at once.

3.3.3 Caching

In the early development stages, we noticed, given the volume of data that comprises the revision history of some articles, and the time complexity of the computations carried over the retrieved data, that it would be desirable to keep already computed results in order to serve those faster if the visualization repeats the request.

In order to minimize this cost and the workload created by our requests on Wikipedia servers, we decided to focus on a cache-based design.

This approach also allows, albeit in an implementation-dependent way, to ensure the reproducibility of test results, although people with such concerns may also set their own MediaWiki instance running with one of the database snapshots made available by the Wikimedia Foundation.

Caches will be used to store the computed the computed results, which comprise both section- and revision-wide content, diffs and statistics, along with the talk page activity.

Caches were chosen to be persistent, providing on-disk storage of data, implemented using Python `shelve` objects, which have the advantage of not needing to be completely loaded in memory — only the requested entries will be loaded from disk. This is an important feature as, for example, caching the content and diffs for the small set of articles chosen in chapter 5 (for user tests) results in a 3 GiB cache[15].

Cache structure: We decided to separate data across two caches, one for content and diffs, and another for statistics, for several reasons:

- Contents and diffs will be requested for single revisions, given their revision id, while statistics will be requested for several revisions at once (as specified by the requested time range), possibly for *the entire* article. These access patterns strongly suggest a revision-centered cache entry approach for the former and an article-centered approach for the latter;
- Statistics are sets of numbers and strings. As said in section 3.1, these are already stored as a Python list. These are not expected to take bigger amounts of space, especially when compared with content and diffs (as an example, the statistics cache corresponding to the 3 GiB user tests cache mentioned above occupies 37 MiB). Thus, separating this data from content and diffs will drastically improve cache efficiency, at least where statistics are concerned.

The content cache will hold a list of content strings (with and without diffs, both split by section), the revision comment, a summary of the number of changes, article-wide (which will be prepended

[14] More correctly, the ordering would depend on the specific *sequence* of requested time ranges.

[15] This cache contains all content and diff entries for the revision history of the entire set of articles, as of May 23, 2012 at 22:12 UTC.

to article-wide requests), and a list with the original ordering of sections, used when concatenating sections to serve article-wide requests.

In the statistics cache, entries are lists containing the statistics (both article- and section-wide) and the talk page activity (timestamps).

Chapter 4

Visualization

In this chapter, we present the design of our visualization (section 4.1), explaining the decisions behind the design, built upon the insights extracted from the surveyed visualization works (as presented in section 2.1.13), with respect to their strengths and weaknesses. The final visualization is then described in detail (section 4.2), where all of its interface elements are introduced and explained. Other architectural choices are discussed later, in section 4.3.

We finish this chapter by presenting an ensemble of case studies where the visualization was used either to explore and compare articles, either to prove or disprove hypotheses regarding some Wikipedia articles (section 4.4).

4.1 Design

All of the works surveyed in section 2.1 focused on a specific visualization type, possibly because it is not easy or not feasible to employ several visualization types together. Therefore, we decided to select one of the several visualization types as the main one to use in our system.

We also decided to use this main visualization at article level, and, from table 4.1, we can see that, of all works which depict metrics for a specific article, the one which covers most of the features is "History flow", a plot-based time series representation focused on content. Thus, we decided that the system would consist of a content-focused plot-based type of visualization.

4.1.1 Plot

Picking "History flow"as a starting point, and inspired by Brandes et al., we suggest a "History flow" and ThemeRiver-like visualization, that is, with a main plot depicting the evolution of content-related aspects along time, with a horizontal timeline, spanning from left to right.

Also inspired by ThemeRiver, which presents data above and below the visualization timeline, we employed a similar approach, dividing the plot in an area above and an area below, which we present in figure 4.1.

	History flow[32, 31]	ThemeRiver[19]	Revert Graph[30]	Brandes et al.[9]	WikiDashboard[29]	WikipediaViz[11]	Wikiswarm[46]
Type	P	P	G	G	P	P	G
Content	✓	✓			✓	✓	
Authorship	✓		✓	✓	✓	✓	✓
Along time	✓	✓		✓	✓	✓	✓
Topic importance		✓					
Author importance	✓		✓	✓	✓	✓	✓
Controversy	✓		✓	✓			

Type key: **P** – plot, **G** – graph.

Table 4.1: Subset of table 2.1, only with visualizations which depict metrics for a specific article.

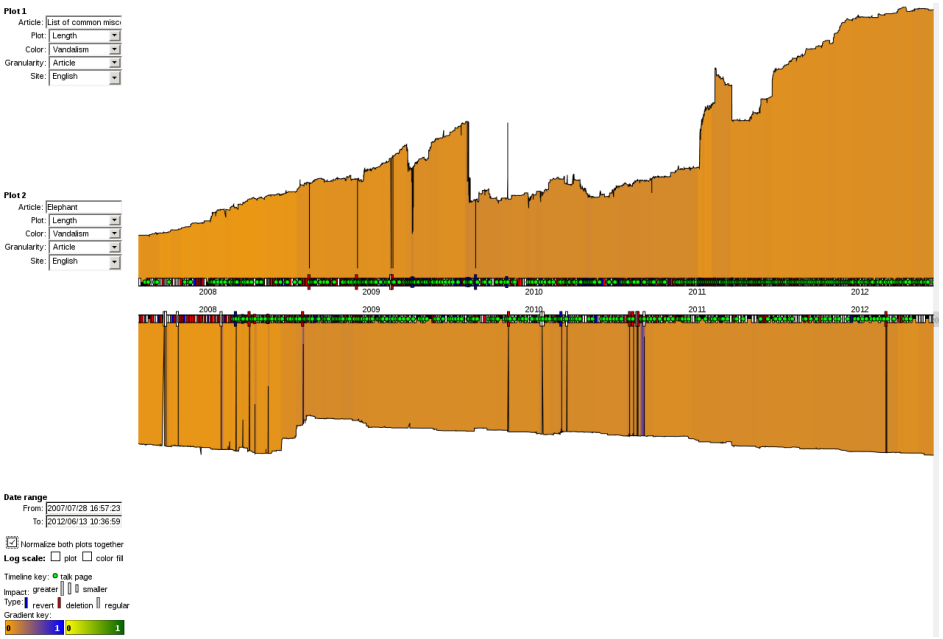


Figure 4.1: The conceived visualization.

Initially, we considered using both areas to plot different versions of the same plot, possibly weighted by different factors, as shown in figure 4.2, but, as restricting the visualization to a single article would seriously limit the power of the visualization, we lifted this restriction and focused instead on presenting two independent plots, drawn over synchronized timelines.

We now describe, in the sections which follow, the several elements that comprise the plot part of the visualization.

Color fill

An attempt was made to graphically convey as much information as possible (unlike the initial prototype of figure 4.2, where most of the information is presented as plain text). As a start, we used

the area under the plot to show additional information through the fill color. This fill is generated according to a gradient, where the numeric value 0 is matched to an orange color and 1 to blue.

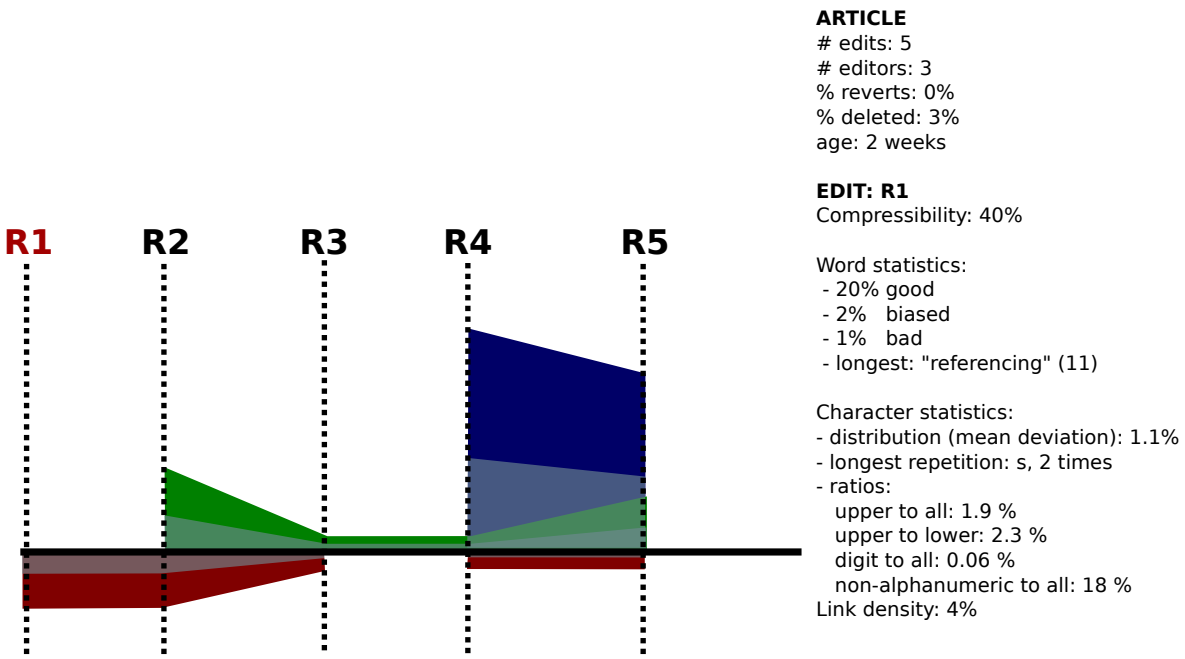


Figure 4.2: One of the first visualization sketches, showing two, differently weighted, versions of the same plot. Revisions, marked by vertical dashed lines, were evenly spread along the horizontal axis, while the vertical axis depicts content length. The plot area is colored by author.

When generating the gradient, redundant gradient stops (with the same color) are handled by removing the middle value each time three consecutive values are found to be identical.

Close stops, which are too close to generate a perceptible effect, are handled in a way similar to superposed timeline markers (discussed in section 4.3), being merged into a single stop, which conveys the maximum value of the merged stops.

Metrics

In an attempt to make the visualization more flexible, we allow the user to pick from a set of metrics to use both as the source for the plot line and for the gradient fill: quality, controversy, vandalism and length, which were defined in section 3.1. We believe this flexibility will allow users to find patterns that would otherwise be lost if the visualization was restricted to a fixed set of metrics.

Authorship

Especially in the view at the section level, described in section 4.1.2, we wanted to provide a "History flow"-like visual feedback of authorship: an additional choice is provided for the color fill, "by author".

A first approach involved the sequential assignment of colors over a continuous range to authors, but this was found to be confusing, especially when analyzing authorship locally, as different authors appearing for the first time on consecutive revisions would get very similar colors, which would eventually be impossible to tell apart by simple visual inspection.

The strategy employed in the final solution orders authors by their edit count for the article being visualized, during the chosen time range, and assigns colors to the twenty most active authors, coloring the others in gray.

This way, we not only make it easier to analyze consecutive contributions, we also enable the user to spot the most frequent contributors and their impact in the article. It allows, likewise, the assessment of the impact of the remaining users, by highlighting their share of the article. We expect that, in popular articles, the share of the less frequent editors will largely outweight that of the twenty most frequent editors, as depicted in figure 4.3.

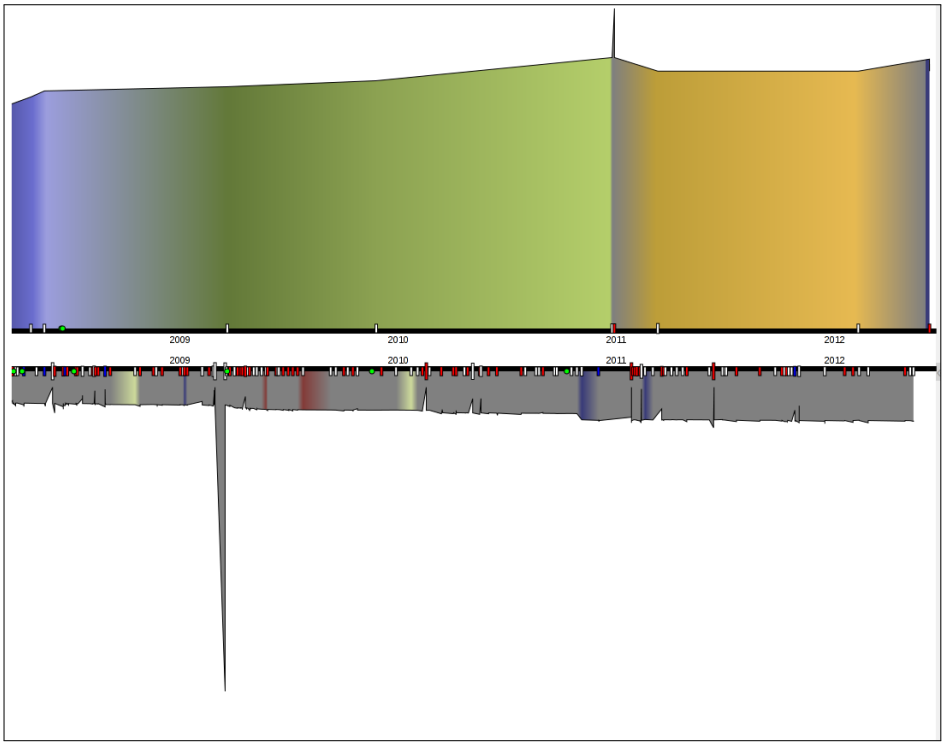


Figure 4.3: Visualization of part of the history of the Portuguese Wikipedia article "Foros de Salvaterra de Magos" (above) and of the English Wikipedia article "42", both plotting length filled by author, where the Portuguese article has almost no less frequent authors, and most of the contributions to the English article were made by less frequent authors.

Optimizing the plot area

In earlier versions, data was plotted with no preprocessing, highlighting not only some issues, which were fixed, but also suggesting some optional features:

Normalization: We initially just normalized the length values, but it soon became obvious that this would need to be done for every metric, in cases as the one shown in figure 4.4 — note how the not normalized visualization (at the left), the bottom plot does not cover the entire plot area. In these cases, the visualization was changed to normalize the incoming data, extending the plot so that its greatest peak uses the entire plot height (as shown at the right of figure 4.4). This feature was later extended, as a result of feedback from user testing, in order to allow the normalization by the

maximum value of both plots, in order to ensure the same metric can be plotted for two articles in a comparable way.

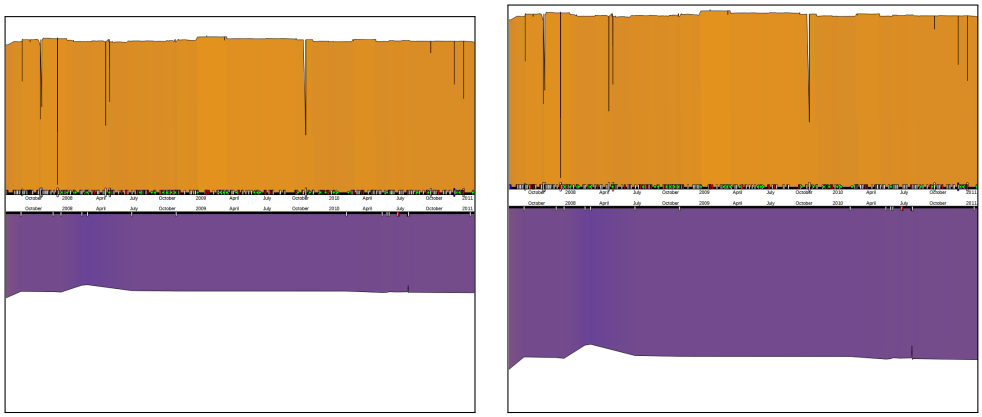


Figure 4.4: Visualization with values as is (left), compared to normalized values (right).

Logarithmic scales: As logarithmic scales are frequently used in science and engineering as a way to smooth exponential growth or otherwise great increases in values, we added an option to enable this manipulation on the metric values. Although the usefulness of this feature depends on the goals of the user and on the specific article being visualized, we believe this will help by highlighting only a specific subset of the metric changes, which the user can then focus on. An example is shown in figure 4.5, where, on the right, a small set of sections from the English Wikipedia "Elephant" article stand out from the rest of the visualization, after enabling the logarithmic gradient mode.

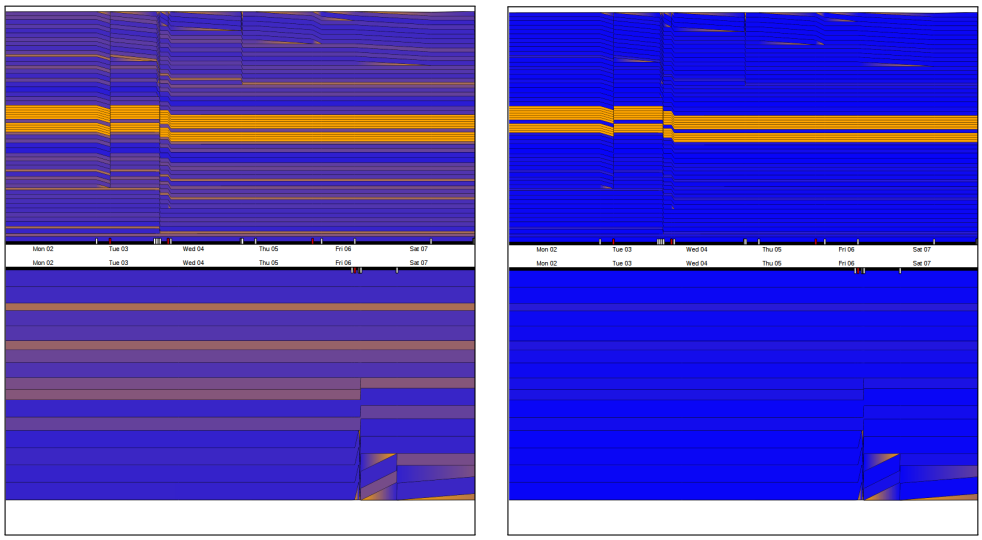


Figure 4.5: Regular visualization (left) and logarithmic gradient fill mode (right).

4.1.2 Granularity

We also wanted to allow the user to find patterns in smaller parts of big articles: the user can set the granularity to be "by section", which shows a set of smaller, stacked plots.

These plots, shown instead of a single plot (granularity "by article"), are rendered using the same settings as the article plot, but their values are computed for a specific section. An example is figure figure 4.5, where granularity was set to "by section".

4.1.3 Timeline

Earlier, we mentioned our visualization follows a time-series approach, thus implying that the plots mentioned in section 4.1.1 are drawn over time.

Although initial prototypes (as, for example, the one shown in figure 4.2) did not take any specific edit spacing rule into account, we later decided to space edits by time.

This is made explicit through the inclusion of a timeline axis, where the time range is conveyed through labels, which present the visualized dates and times.

We also wanted to convey some information on edit magnitude, and highlight reverts and deletions in some way. As edits occur along time, we opted by showing that information directly in the timeline.

This gives the user a way to assess the edit rate, which translates into timeline marker density. While this was already enough to assess the popularity of the article among wikipedians, we also expected talk page activity to relate to popularity in a similar way.

Considering that, in some cases, high-traffic articles get temporarily protected against changes from some sets of users (for example, unregistered and recently registered, but possibly covering a larger part of the community), an increase in popularity or controversy may end up translating into a talk page activity increase instead of an article activity increase.

To address this concern, and given how adequate the timeline was to convey the edit information, we decided to use the same approach to present talk page activity, using differently shaped markers (figure 4.6).

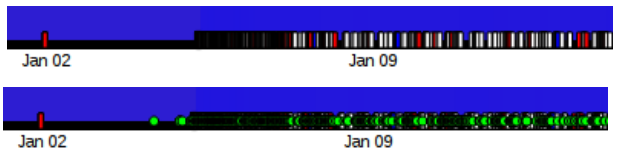


Figure 4.6: Timeline, with article activity only (above) and also including talk page activity (below).

Interaction

Although our initial prototypes (such as figure 4.7) already had input boxes to configure the time range, we have also added mouse-based ways to change this range: Users can *zoom* the plots, by clicking, dragging and releasing the mouse pointer over a plot area, and *pan* the visualization by dragging the area between the two timelines.

4.1.4 Interaction Feedback

Some operations depend on external data, which may not be available or may take a long time to process. In those cases, we provide users with some feedback.

In the initial design, denoted in figure 4.7, articles were specified through a simple text input field. There was no way the user could be warned beforehand if the article did not exist, or informed about articles with similar names.

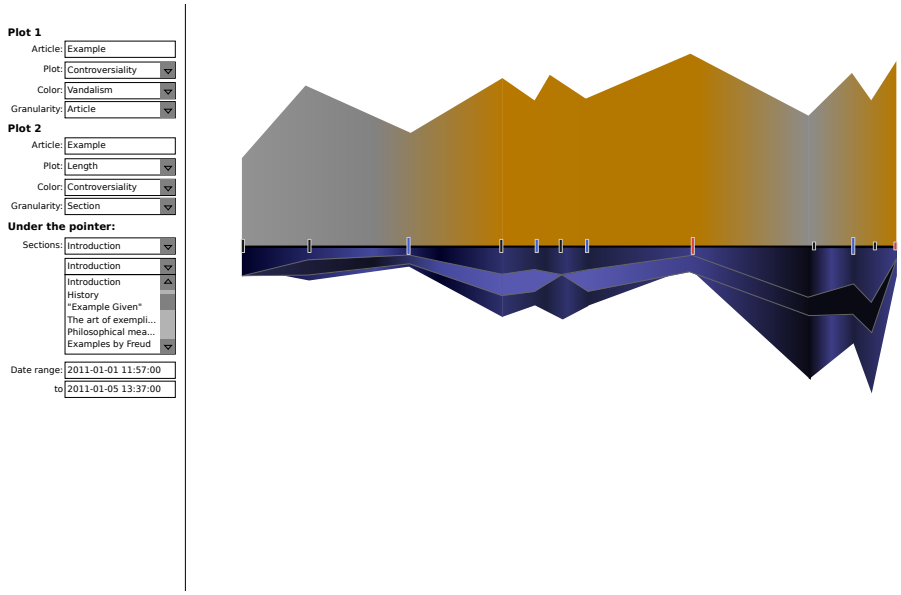


Figure 4.7: Early draft of the visualization, already presenting the independent plots.

This was later solved using the MediaWiki API search suggestion feature to provide feedback by showing a list of suggestions below the input box, and highlighting the input field with a red background when there is no article with the entered name.

Figure 4.8: New real-time AJAX article search: a search for “Rui” showing some results (left), and a search for “NCSA Mosaic” with no results (right).

Likewise, initial prototypes had no way to convey feedback on article processing, which may take some minutes for more active articles. Newer prototypes now show a small progress message, as in figure 4.9.

Figure 4.9: Progress message.

4.1.5 Tooltip

Some possibly insightful information which was not associated to any graphical entity was then presented through a table in a textual tooltip.

This table, shown in figure 4.10, also includes the numeric values of the metrics used to draw the plots.

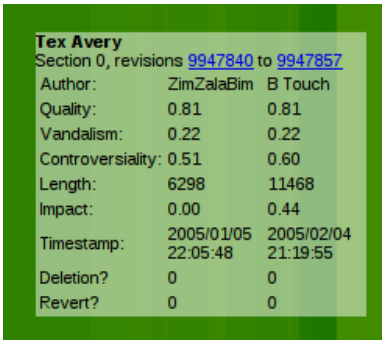


Figure 4.10: Tooltip being shown in the visualization of the English Wikipedia article on "Tex Avery".

4.1.6 Content Pane

As we wanted users to be able to investigate the revision content without leaving the visualization, we added a togglable content pane.

This pane, depicted in figure 4.11, shows the content of the revision under the pointer, or its differences with respect to the previous revision.

This addition should greatly improve the capabilities of the visualization, as, while the metric values only allow users to infer hypotheses about the article under study, the content, along with its changes, will enable users to assess the correctness of their hypotheses.

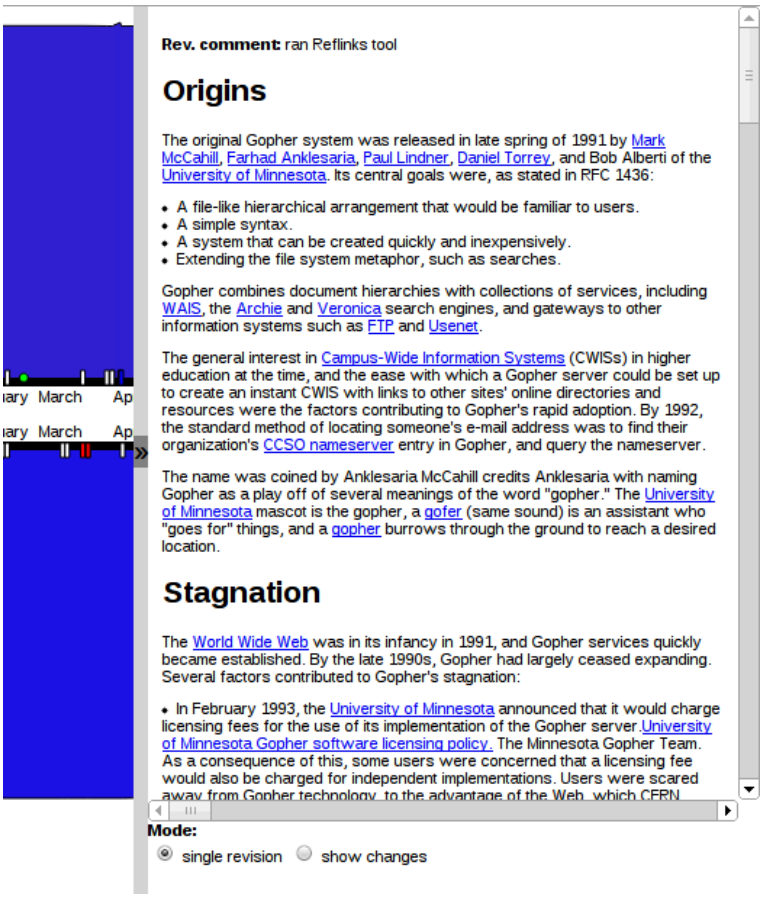


Figure 4.11: Visualization with the content pane enabled, showing content from the English Wikipedia article on "Gopher (protocol)".

4.1.7 Comparison With Related Work

Our decisions have been largely influenced by "History flow"[32, 31] and "ThemeRiver"[19], leading to a similar overall approach, albeit with some important differences.

Like these works, our visualization conveys time-series data, which is discrete by nature, but is represented through a continuous plot. Although the lines connecting data points do not have any meaning, these are expected to ease the use and interpretation of the visualization.

Both "History flow" and "ThemeRiver" do not allow users to change the depicted metrics, locking users to the specific metrics chosen by the authors of the visualization. In this area, our visualization is more flexible: it allows users to switch between several metrics, enabling users to extract more insights from the data.

On the other hand, "History flow" offers users the ability to choose between two different revision spacing behaviors: revisions can either be spaced by time or evenly distributed along the axis, while our work only offers the first approach.

In our system, the most detailed granularity level is the section level, where plots for each section are stacked on top of each other. Like "History flow", we allow users to keep track of a specific section along time, by highlighting a section.

"History flow" also offers a deeper level, the contribution level, where each segment corresponds to text contributed by the same author, explicitly conveying authorship information that is lost is our visualization. As part of this deeper granularity level, "History flow" also enables users introduces to filter out some of the content, by choosing a specific author. Contributions by other authors are then hidden, allowing the user to focus on a specific author.

"ThemeRiver" requires human input in order to group the depicted topics. It also uses colors to represent the different topics, an approach that does not scale well for corpora with a large number of topics. None of these limitations are present in our work, which relies on the automatic processing of articles and does not attempt to identify topics.

Compared with the remaining surveyed works, our system differs significantly from Revert Graph[30], WikiViz[18], ClusterBall[17] and Omnipedia[6] as these works do not portrait the evolution of metrics through time, and, to a lesser extent, from Brandes et al.[9], Holloway et al.[20] and Wikiswarm[46], as those works either do not integrate the time series information with the main visualization or in a single screen.

Both WikiDashboard[29] and WikipediaViz[11] focus on a time-series approach, but these visualizations are dominated by the content of a single revision of the article, while our system accommodates most of the available screen area to the display of metrics themselves. As our system focuses on the time series, it makes it easier for users to navigate through the content of different revisions.

While iChase[26] conveys several aggregated information on a screen with different visualization elements, we opted by a single major element (the plot). We believe that, while iChase might help users extracting intricated patters, using its different visualization and comparison techniques, our system makes it easier for users to extract patterns in general.

It should also be noted that iChase focuses on interarticle analysis. While our work also allows interarticle comparisons, our visualization elements are restricted to a single article. On one hand, our tool will not enable users to analyze the overall trends of an entire Wikipedia Project. But, on the other hand, it conveys information on a single article in a simpler, clearer way.

4.2 User Interaction

The visualization is made of three parts, one of which comprises two plots, used to depict the metrics extracted from two articles. These parts, except for the content pane, which is not shown by default, are visible in figure 4.12 and are as follows:

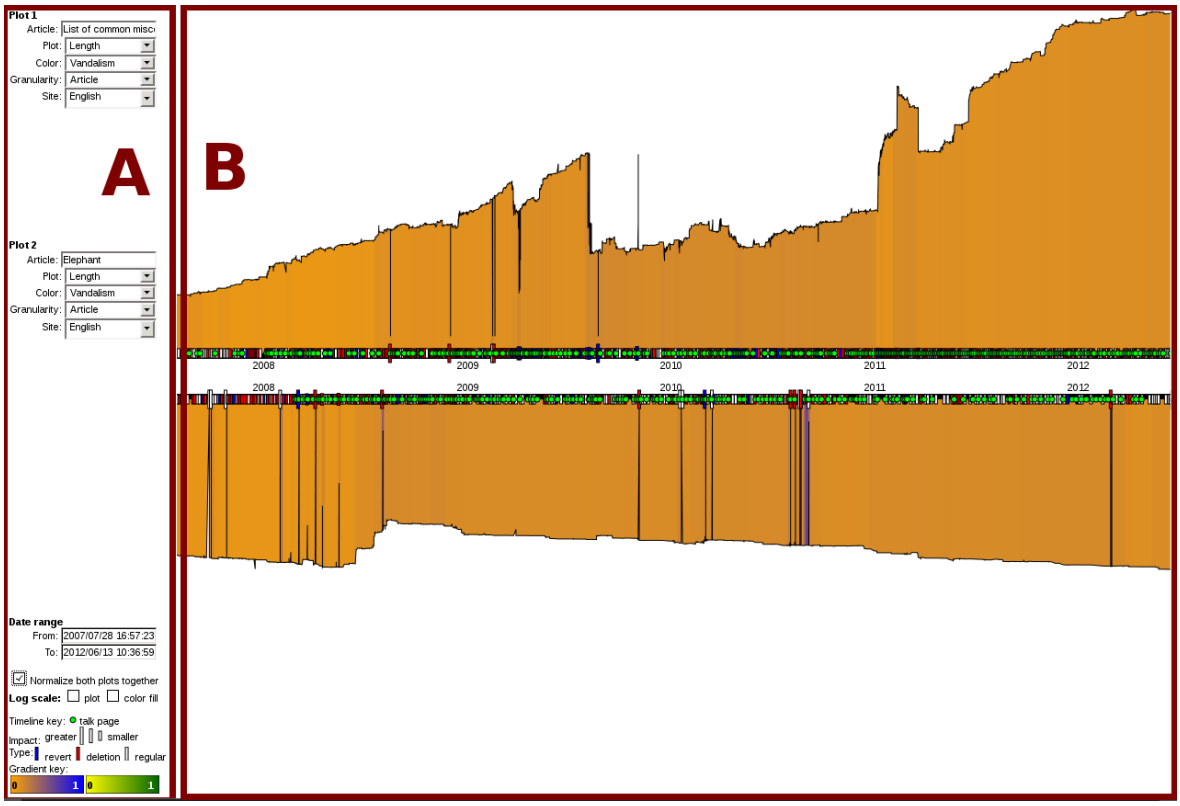


Figure 4.12: The conceived visualization: A) the options pane; B) the plot area.

The Option Pane: Located at the left side of the screen (figure 4.12A, figure 4.13), provides options for the whole visualization (figure 4.13C), options for each of the plots (figure 4.13A and B) and a key explaining the meaning of some of the graphical elements used to convey information (timeline markers, regular gradients and highlighted gradients[1], figure 4.13D).

Figure 4.13: The options pane: A) Upper plot options; B) Lower plot options; C) General options; D) Visualization key.

[1]When the mouse pointer is over a plot area, its colors are changed to the highlighted gradient, in order to distinguish it from other plots in the visualization.

For each plot, users are given an option pane (figure 4.13A and B, shown in detail in figure 4.14) allowing them to

- Change the article under analysis, through an "Article" textbox (figure 4.14A), which employs a real-time AJAX search on the Wikipedia database, whose results are shown in a list which will appear below the textbox;
- Choose the metric used in the plot (figure 4.14B), one of "controversy", "vandalism", "quality" and "length".
- Choose the metric used to fill the plot (figure 4.14C), for which all the four metrics above are available, along with a fifth one, "by author", which colors revisions by their author (attributing colors to the twenty most common editors, and painting the remaining ones in gray).
- Change the granularity of the visualization (figure 4.14D), between article-level and section-level.
- Select a Wikipedia instance to access (figure 4.14E).

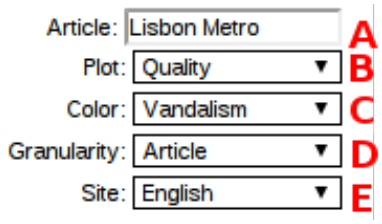


Figure 4.14: The plot options pane.

Users can also access general options (figure 4.15) in the bottom left corner, concerning the visualization timespan ("Date range", figure 4.15A), the normalization strategy, which can be set to normalize to the maximum of *both* plots (figure 4.15B), and toggling logarithmic scales ("Log scale", figure 4.15C), both for the plot metric and for the filling color.

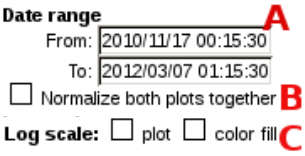


Figure 4.15: The general options pane.

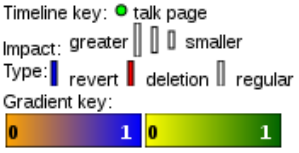


Figure 4.16: The visualization key.

A **key** (figure 4.16) is also available, explaining the colors and symbols used in several parts of the visualization.

The plots: The central part of the visualization, lying on its center and extending to its right, are two plots along horizontal timelines (figure 4.12B), which encode information about edits, marked using small rectangles and color coded according to their kind (blue for reverts, red for deletions and white for regular edits, as listed in figure 4.16), and edits to the corresponding talk page, marked with green circles. The impact of page edits is conveyed through the height of their markers, which is proportional to that impact. A screenshot of a plot is shown in figure 4.17.

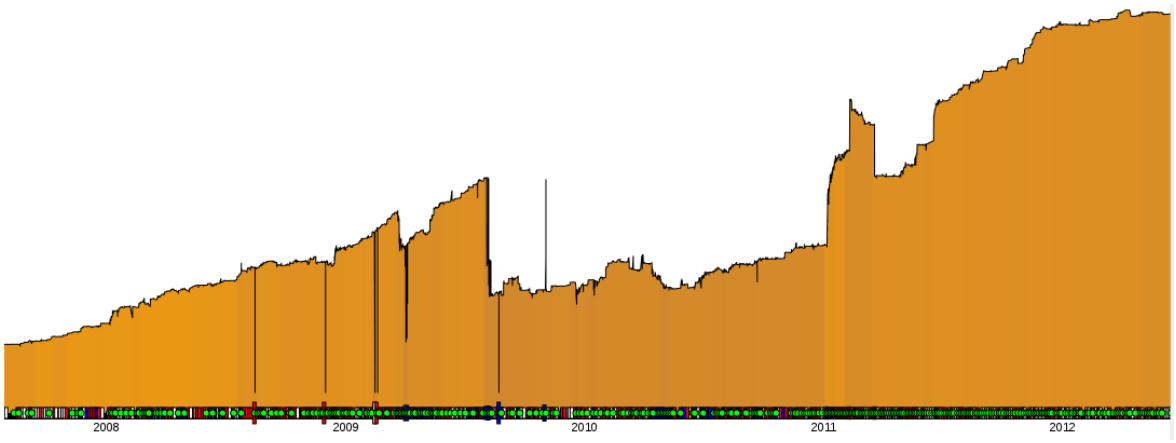


Figure 4.17: A plot.

When the mouse pointer is over the plot, some textual information about the segment under the pointer is shown in a tooltip near the pointer (figure 4.10). Clicking with the left mouse button inside the plot locks this information for further analysis, which can be unlocked by left clicking the plot again.

This information includes the covered date ranges, the involved section (if the granularity is "by Section"), and the involved editors.

Content: Right to the plot area, there is a pane, closed by default (figure 4.18), but togglable using a small button on the right side of the screen (figure 4.19), which shows, according to the selected mode, either the revision content or the differences between the two consecutive revisions under the pointer.

This pane can be resized, by dragging and dropping its left edge.

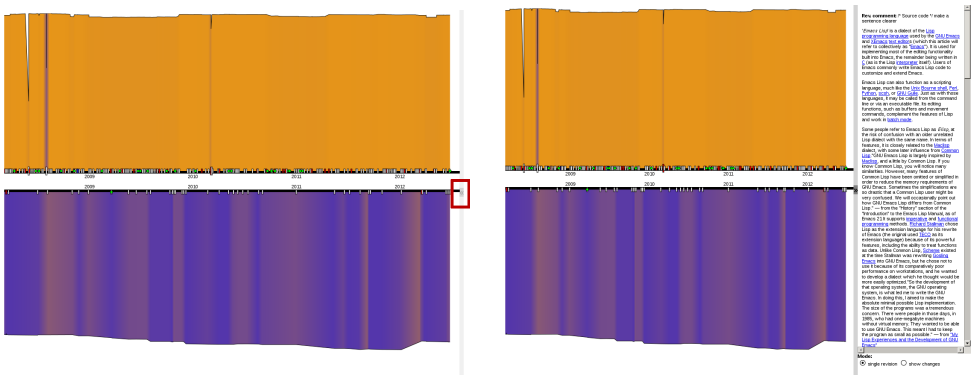


Figure 4.18: On the left, the visualization, with the content pane toggle button highlighted. Clicking in the button opens the content pane, as shown in the right.

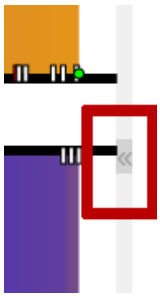


Figure 4.19: The content pane toggle button.

Interaction: Plots are updated in real time: after editing some option in the textboxes (article name and dates), the update is triggered by leaving the field or by hitting the "Enter" or "Return" keyboard keys.

The content pane is updated when moving the mouse pointer over the plot, unless the tooltip has been locked.

The plot can be *zoomed into* by selecting (click, drag and raise the left mouse button) a plot area. Zoom out is achieved by right-clicking the plot. It can also be *panned* to the left or right by dragging and dropping the area between the two timelines.

4.3 Implementation

For the implementation of this visualization, we decided to build upon d3.js[8], a JavaScript visualization and graphics library, and jQuery, a generic library of JavaScript tools, together with the asynchronous retrieval of the article-related data through HTTP requests to the component described in chapter 3, thus being an AJAX[2] application running on a JavaScript and SVG-capable web browser[3].

[2]Asynchronous JavaScript And XML, a name used to denote interactive applications which rely on JavaScript computation of asynchronously-requested data. Although nowadays data is usually in formats other than XML, the name is still used to describe the whole family of interactive applications, even ones built upon JSON or HTML.

[3]Google's distribution of the Chromium web browser, "Google Chrome", was used during the development and testing of the system, given its focus on being a platform for rich JavaScript-based applications, but any d3.js and jQuery-compatible browser should be able to run the system, albeit with possible speed differences.

Other than the metrics assessment component, the visualization also interacts directly with the MediaWiki API (figure 4.20), in order to retrieve the search suggestions list, when the user changes the article name text box.

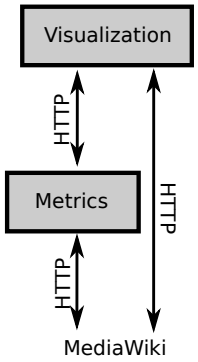


Figure 4.20: High-level architecture of the system, focusing on the interactions between the visualization and other entities.

Plotting data: The visualization operates by issuing asynchronous requests for the data, which, when served, are preprocessed and used to build the `d3.js`-based plot, which involves the generation of SVG graphics from the metrics, a data flow we present in figure 4.21.

Early prototypes were mostly based on the `d3.js` "line" example[4], which was extended in order to support gradients as the fill style and to allow stacked plots.

Plots are generated by creating a plot area from the datapoints, which is then filled with a generated gradient. These steps are executed for each plot, as there may be several plots, if the granularity level is set to "by section". Finally, the plot timeline is updated and its markers are generated. This processing is completely carried out inside the response handler, as it builds the elements of the visualization.

When drawing the plot line and the boundaries of plot areas, numeric metrics (except length) are provided in the range $[0, 1]$, being processed in order to normalize the values so that the highest value becomes 1. Lengths are preprocessed in a similar way, ensuring that the final value fits in the $[0, 1]$ range. In the plot, where height is concerned, the range is then mapped so that $1 \cdot n$ (for n stacked plots) corresponds to the height of the screen area available for that visualization.

Plot lines are also colored black, in order to ensure the user is able to tell sections apart when the metric used in their fill presents the same value (resulting in an identical fill).

Gradients map values in the same way, using another `d3.js` "scale" (an object which maps values across two ranges), this time using a *color interpolator*, where 0 corresponds to orange (or yellow, in highlighted gradients) and 1 to blue (green, in highlighted gradients). Values between 0 and 1 are mapped to corresponding shades of orange and blue (or of yellow and green, for highlighted gradients).

Timestamps are passed to those objects in order to define the points of plot lines, boundaries of plot areas and gradient stops. Once again, `d3.js` scales are used to map timestamps to the visualization window.

[4] https://github.com/mbostock/d3/tree/master/examples/line

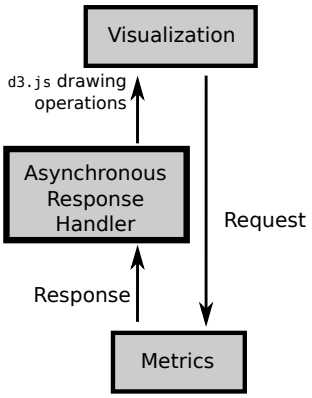


Figure 4.21: Depiction of the process through which the visualization is updated: an asynchronous request is made by the visualization to the metrics component, which is later handled by a response handler, which will rebuild the `d3.js`-based visualization.

Optimization: Spacing data by time has the advantage of rendering the horizontal axis useful, conveying additional information, while also depicting their dispersion over time.

But this dispersion results in the superposition of close edits, leading to cluster of markers where only one marker will be visible. This not only results in the computational overload of computing and drawing timeline markers that will be rendered useless as they are covered by other markers, but also in the undesired visual effect of the superposition: the information conveyed by the covered markers is lost.

In order to address the problem of marks being too close together, these are merged into a single marker, whose height is the highest of all heights, thus conveying the strongest edit impact among the merged markers, and which is colored as revert or deletion if and only if at least one of the merged markers are marked as such.

When rendering the markers, priority is given to reverts: if a revert is merged with a deletion, the final marker will be colored as a revert, just like a single edition would be if it were marked as both a deletion and a revert. This follows the belief that, if both events took place, then the most important one is the revert, as it implies disagreement with a previous edit, while a deletion may be triggered by something other than disagreement (for example, removing no longer relevant content in an article on a current event).

With this optimization, both the gradient fill and the timeline markers are simplified, and the final number of gradient stops and markers is constrained by the available screen pixel width.

Customization: One of our goals, when designing the visualization, was to avoid any ties with specific MediaWiki instances. As far as an instance provides the API features used both by the visualization and by the metrics assessment component, it can be used with this visualization.

In order to allow the easy tweaking of the list of available servers, the server parameters are specified once, in a single call to a JavaScript function, where the parameters are given, along with a descriptive name that is used to list the MediaWiki instance in the site selection drop-down box provided by the user interface.

This configuration is done only once, in the visualization, even if both the visualization and the metrics assessment component require these parameters to identify the MediaWiki API endpoint (figure 4.22).

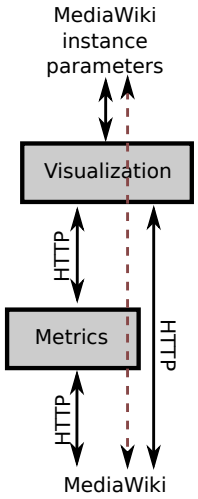


Figure 4.22: High-level architecture of the system, focusing on the handling of the parameters which identify the specific MediaWiki instance in use.

By default, the code installs the 40 largest Wikipedia editions: English, Deutsch, Français, Nederlands, Italiano, Polski, Español, Русский, 日本語, Português, 中文, Svenska, Tiếng Việt, Українська, Català, Norsk (Bokmål), Suomi, Čeština, Magyar, 한국어, فارسی, Bahasa Indonesia, Türkçe, Română, العربية, Slovenčina, Esperanto, Dansk, Српски / Srpski, Lietuvių, Euskara, Bahasa Melayu, Slovenščina, עברית, Български, Қазақша, Volapük, Winaray, Hrvatski and हिन्दी.

4.4 Finding Patterns in Wikipedia

After implementing both components, we then used the system to explore the revision history of several articles from some of the available editions of Wikipedia.

4.4.1 Article Patterns

We first started by looking for article-level patterns, either by using only one of the plots, or by plotting the same article on both plots.

Vandalism and length: When exploring articles expected to be highly vandalized, it became evident that length as a metric, along with the employed normalization, made some acts of vandalism evident, as these translate into sudden, abrupt length changes.

These changes may involve the addition of a large amount of unrelated or undesirable content (defined as "phony copy" by Viégas et al.[32], and which we could refer to as *mass insertions*) or the sudden removal of a great part of the article content (defined as "mass deletion" by Viégas et al.). This pattern is depicted in figure 4.23, and its nature can be assessed using the content pane: some user added numerous repetitions of the expression "ITSTHEMEANINGOFLIFE" to the first paragraph of the English Wikipedia article on the year "42" of the Julian calendar (figure 4.24).

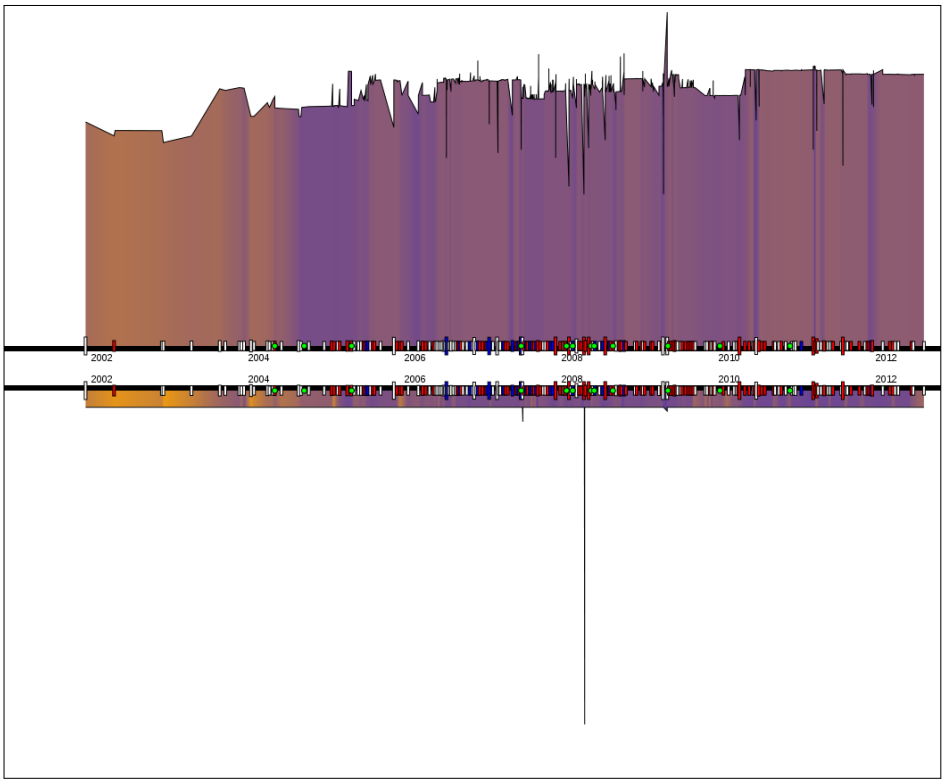


Figure 4.23: Visualization of the English Wikipedia article “42”, where the lower plot is set to use length as the plot metric and controversy as color fill, with the upper plot set to use quality and fill by vandalism.

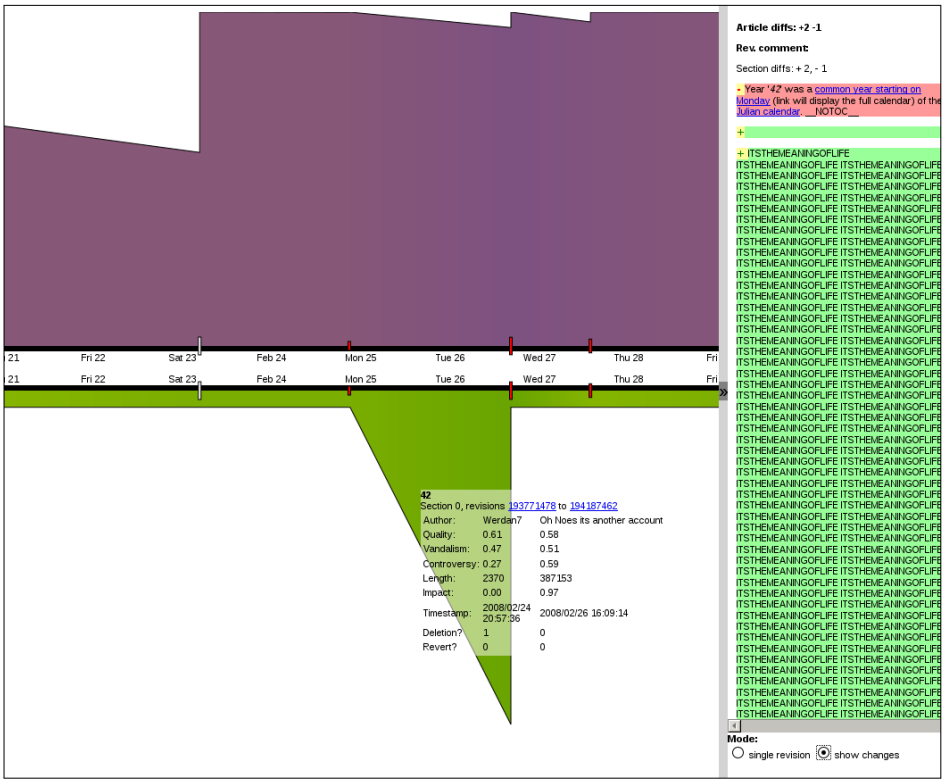


Figure 4.24: Visualization configured as in figure 4.23, focused on a *mass insertion* of undesired content, which can be seen in the content pane.

Mass deletions were also found to be easy to spot using length as the plot metric, either when seeing a "big picture" visualization of the article over several years (figure 4.25), or by focusing on a narrower range (figure 4.26).

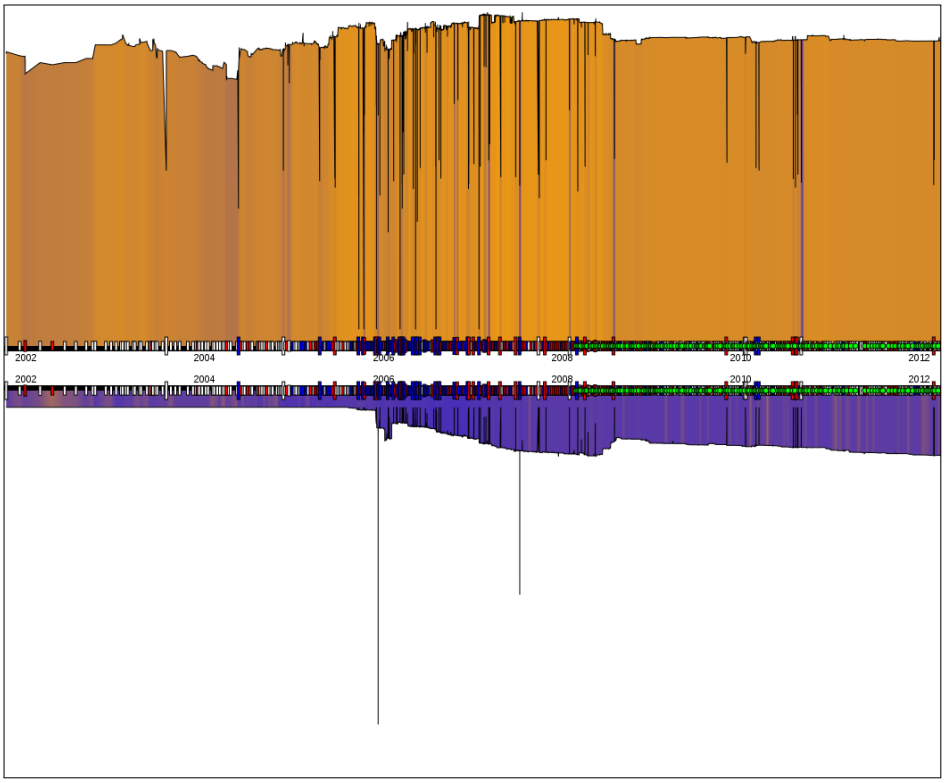


Figure 4.25: Visualization of the English Wikipedia article "Elephants", where *mass deletions* can be spotted as the cause of sudden changes in the lower plot line (length, filled by controversy). The upper plot is configured to use quality and fill by vandalism.

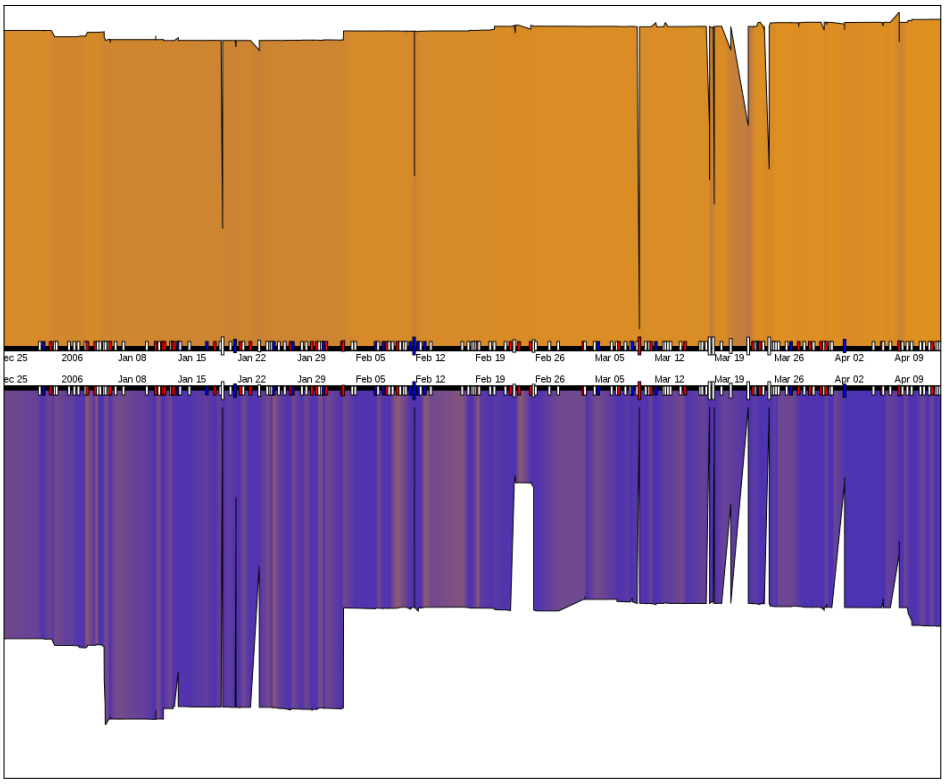


Figure 4.26: Visualization of figure 4.25, focused on a shorter timespan where some *mass deletions* become more evident.

Popularity and talk page activity: The English Wikipedia article on the Republic of Kosovo was protected on 27 February 2011, following an edit war on whether to redirect the page to the article on "Kosovo" or to have a separate article on the "Republic of Kosovo"[5]. As the article was placed under protection, the sudden interest in the article did not translate into a heavy increase of the edit rate. It was instead observed as increased talk page activity, as shown in figure 4.27, where a "zigzag" pattern corresponding to the edit war can also be observed (another of the patterns identified by Viégas et al.).

[5]As of June 2012, the two articles coexist separately, along with an article on the "Autonomous Province of Kosovo and Metohija", "Kosovo" being an article on the geographic region itself and the other two articles on the entities claiming sovereignty over the geographic region.

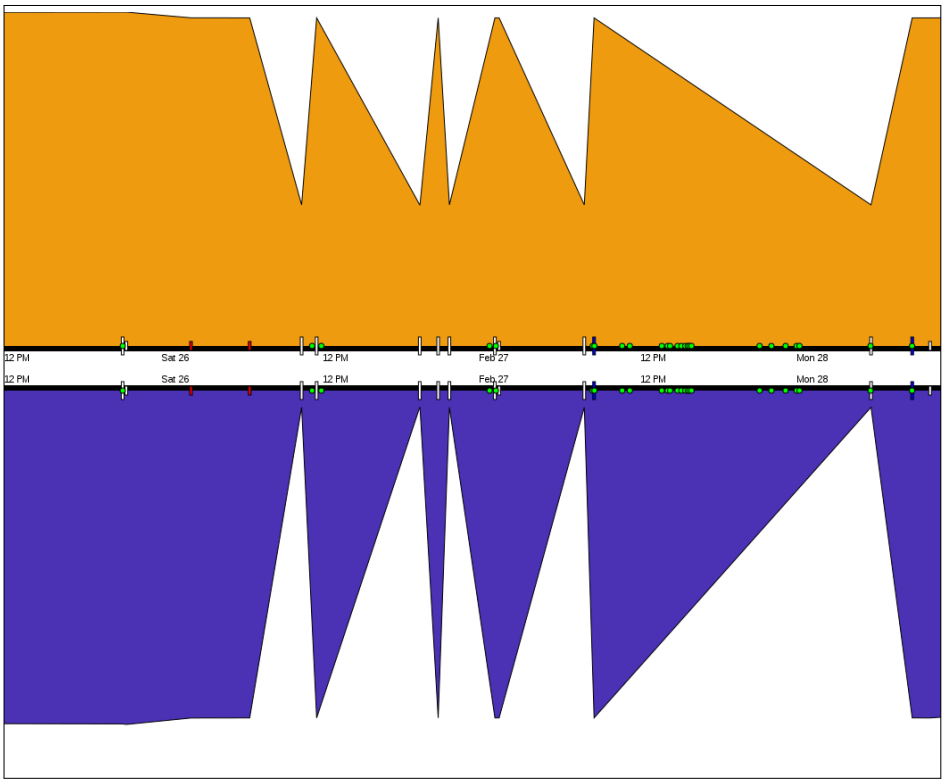


Figure 4.27: Visualization of the English Wikipedia article on the "Republic of Kosovo": an edit war regarding the split from "Kosovo" and increased talk page activity. (Upper plot: quality, filled by vandalism; lower plot: length, filled by controversy.)

Visualizing popularity: With the visualizations, it is possible to assess the article popularity by looking at the timeline (although length changes may also be a sign of increased popularity). In the case of the article "List of common misconceptions", it is possible, by looking at the big picture, to discover the article became popular in 2006, and has seen an increased edit rate ever since (figure 4.28), but, by looking closer, it is also possible to see periods when the article became even more popular — the visualization enables us to find that there was a sudden activity increase during January 2012 (figure 4.29), possibly due to the article having been linked from several media (Randall Munroe drew a comic on the article[23] and Cory Doctorow wrote an article on it at BoingBoing[15]).

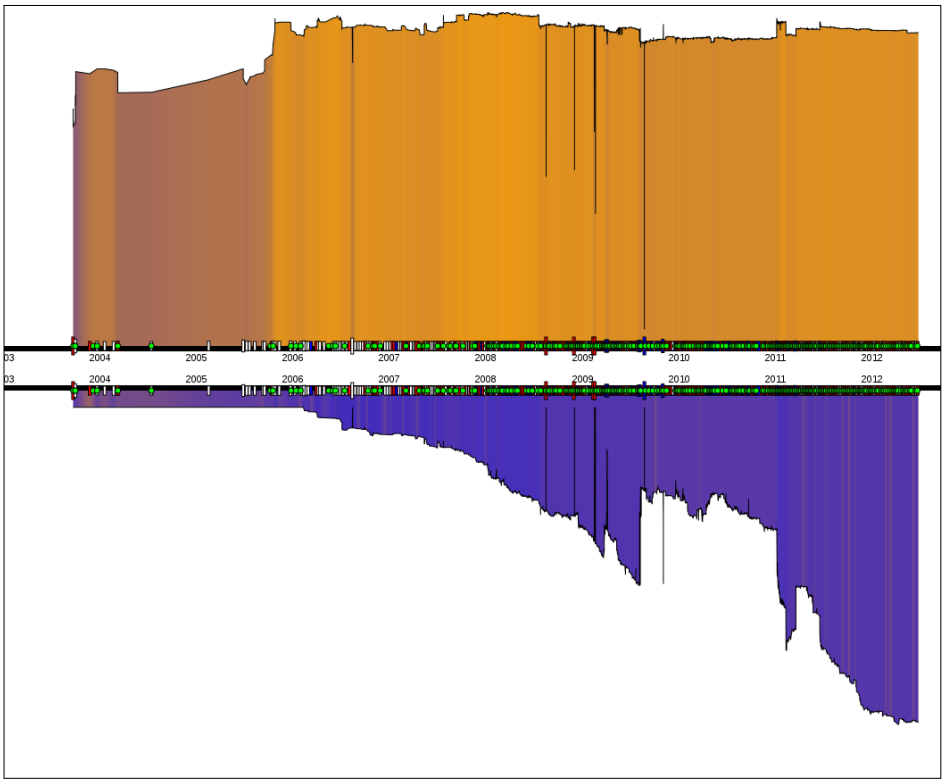


Figure 4.28: Visualization of the entire history of the English Wikipedia article on the "List of common misconceptions". (Upper plot: quality, filled by vandalism; lower plot: length, filled by controversy.)

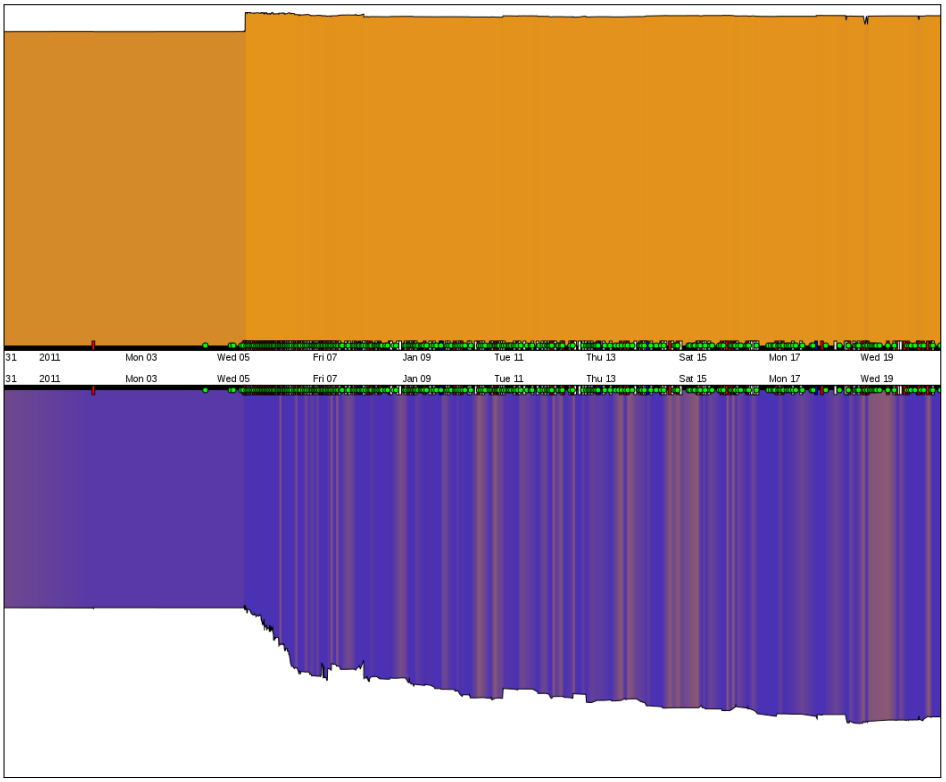


Figure 4.29: Visualization of figure 4.28, focused on the time range from 31 December 2010 to 21 January 2011, showing a sudden popularity increase on January 5.

It is also likely that the visualization helps assessing whether an article is highly vandalized or simply more popular. One example is the comparison of "List of common misconceptions" with "Chicken", an article which has been targeted with vandalism since Ryan North suggested "Instead of vandalizing Wikipedia in general, we all just vandalize the chicken article."[24] (figure 4.30): the "List of common misconceptions" article, despite having a high edit rate, does not show as many reverts and deletions as the article on "Chicken". In this case, the "Chicken" article also exhibits visible fluctuations (blue strips, corresponding to higher values) in the "Vandalism" metric, while the fill color of "List of common misconceptions" stays the same across the visualization.

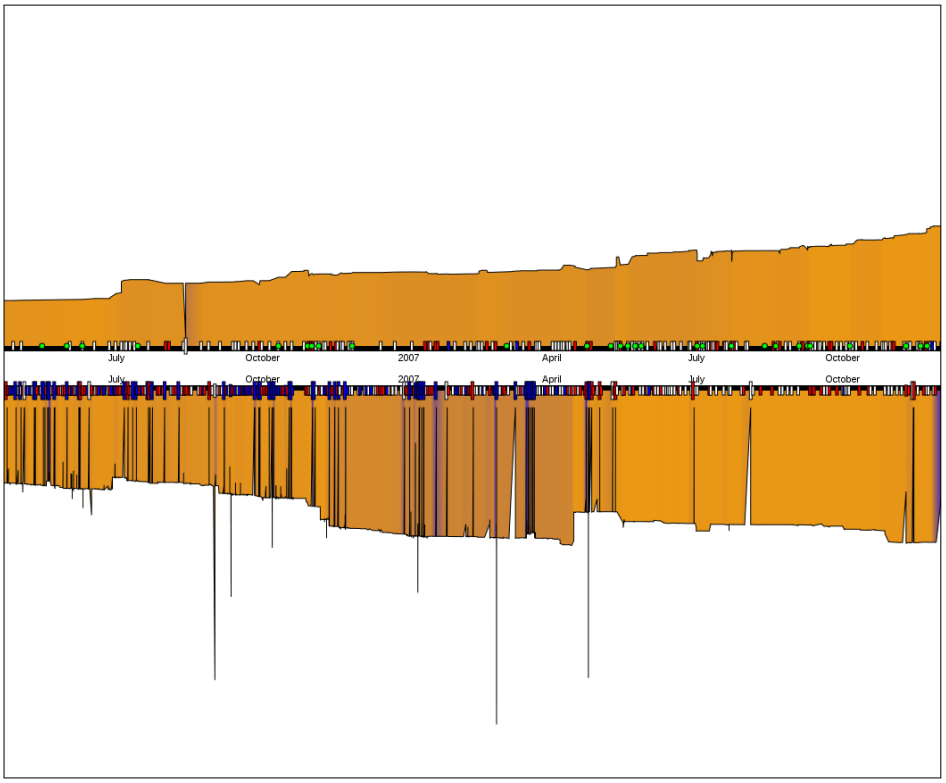


Figure 4.30: Visualization of the English Wikipedia article on the "List of common misconceptions" (above), together with the article "Chicken" from the same encyclopedia (below). Both plots are set to use length and fill by vandalism.

Focusing on the visualization of "Chicken" from September 4 to October 20, 2011, we can see how the several metrics are able to depict the article revision 452274414[6] (where user "Wearefictional" replaced the article content with "The chicken Delicious"), which translates into a lower quality value, a shorter length and a higher vandalism value, as depicted in figure 4.31, where the vandalism revision clearly stands out in the vandalism fill gradient (marked in blue against a uniform orange fill).

[6] `http://en.wikipedia.org/w/index.php?oldid=452274414`

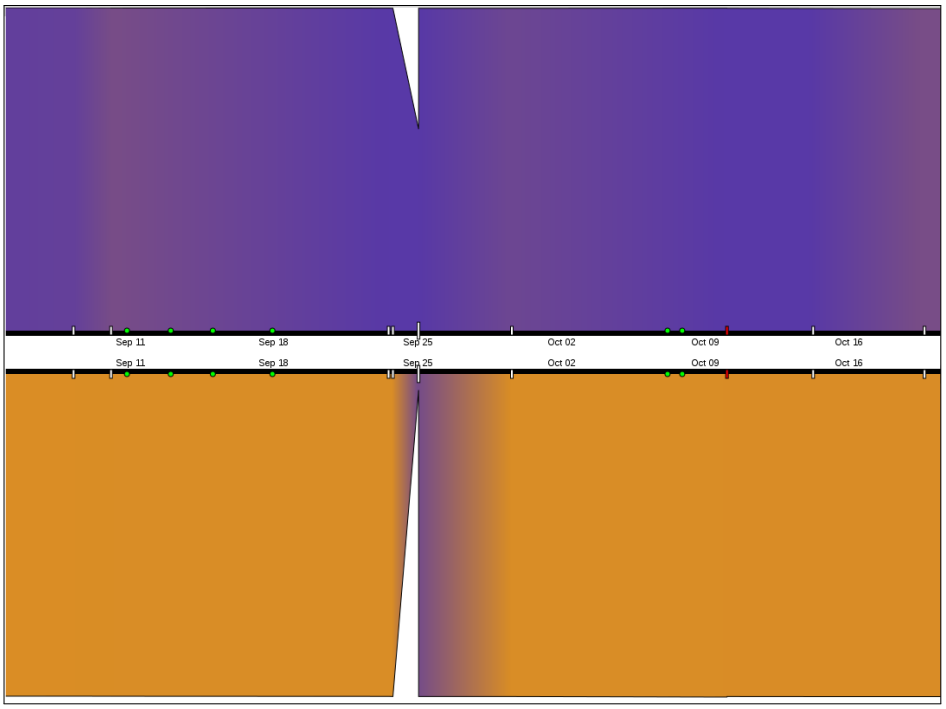


Figure 4.31: Visualization of the English Wikipedia article on "Chicken". The upper visualization is set to show quality and fill by controversy, while the lower one is set to plot length and fill by vandalism.

4.4.2 Comparing Articles

A key feature in our visualization is the ability to compare independent articles on the same screen, either articles known to be on related topics, or completely distinct articles.

Computer Programming Languages: There are plenty of programming languages, some of which are popular only for a short period of time, others that continue being used for many years. By exploring articles on programming languages, we found that:

- "C (programming language)", the article on the well-established programming language that arose from the development of the UNIX operating system, is much more popular (and vandalized) than the article on "Emacs Lisp", the Lisp dialect used in the Emacs text editor (figure 4.32). "Emacs Lisp" is also not as popular as "Lisp (programming language)" (figure 4.33), the article on the family of languages of which Emacs Lisp is a member.

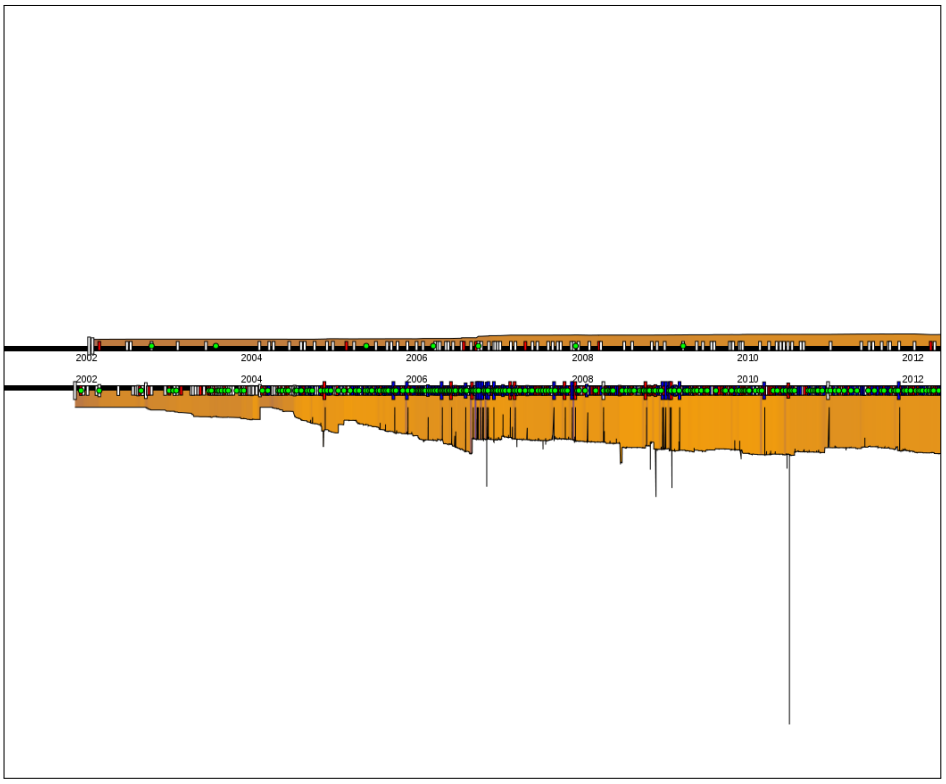


Figure 4.32: Visualization of the English Wikipedia articles on Programming Languages: "Emacs Lisp" (above) and "C (programming language)" (below). Both plots are set to use length and fill by vandalism.

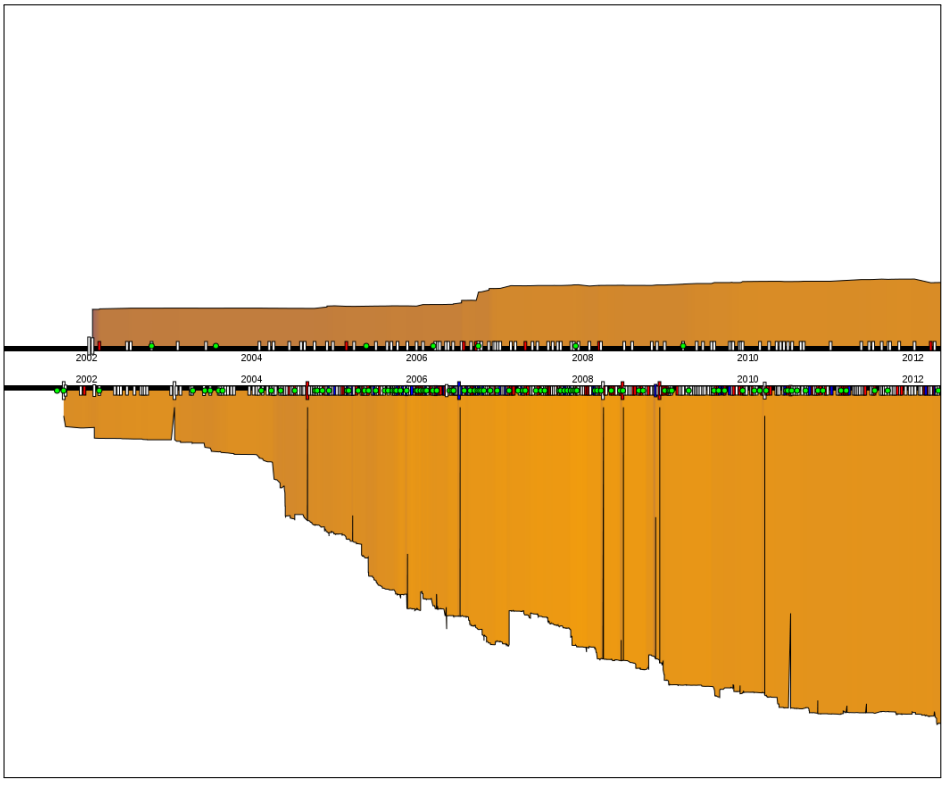


Figure 4.33: Visualization of the English Wikipedia articles on Programming Languages: "Emacs Lisp" (above) and "Lisp (programming language)" (below). Both plots are set to use length and fill by vandalism.

- Lisp and C are more popular languages, with a similar long-term activity pattern, although the article on C appears to be more vandalized (judging from the number of reverts) and has more talk page activity (figure 4.34).
- The same observation can be made for "C++ (programming language)" and "BASIC", where both articles are significantly active, although the former seems to be more frequently vandalized. This distinctions grows especially stronger in the earlier days of the article, when the article on BASIC was significantly less popular than nowadays, unlike the article on C++, which shows an (approximately) steady popularity level. As BASIC appeared in 1964, this distinction between the articles may imply a lack of representativity of the BASIC community in the earlier days of Wikipedia, contrasted with a much more representative universe of editors interested in C++.

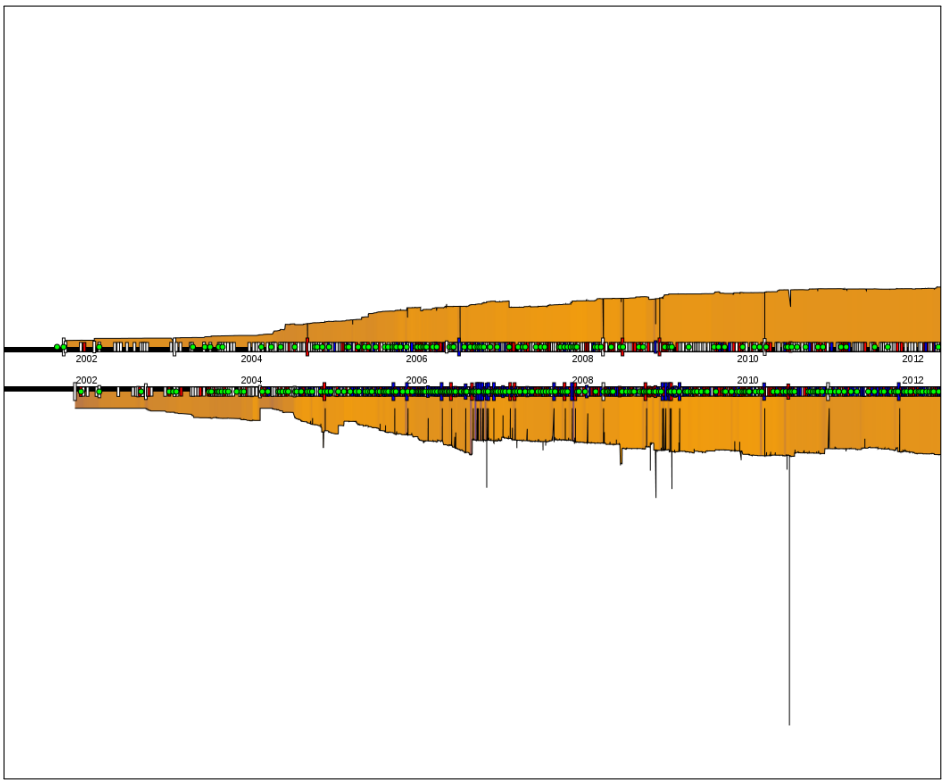


Figure 4.34: Visualization of the English Wikipedia articles on Programming Languages: "Lisp (programming language)" (above) and "C (programming language)" (below). Both plots are set to use length and fill by vandalism.

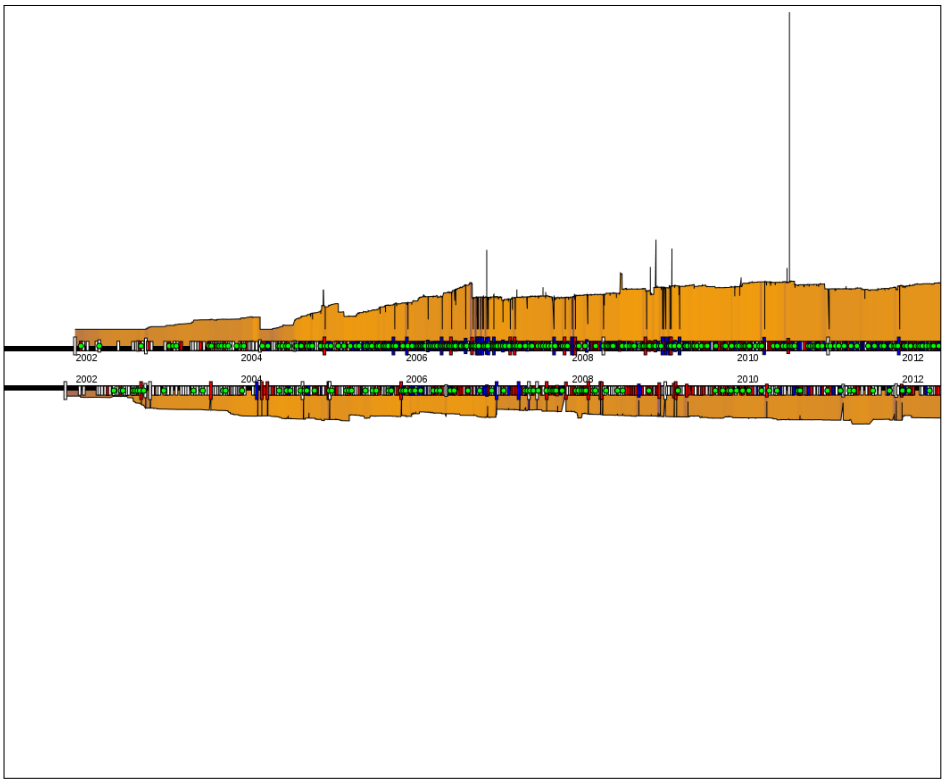


Figure 4.35: Visualization of the English Wikipedia articles on Programming Languages: "C++ (programming language)" (above) and "BASIC" (below). Both plots are set to use length and fill by vandalism.

The findings extracted from the timelines are summarized in figure 4.36, where the lack of popularity of Emacs Lisp stands out compared to C++, BASIC, C and Lisp. It is also possible to observe that the articles on C++ and C have much more reverts than the others.

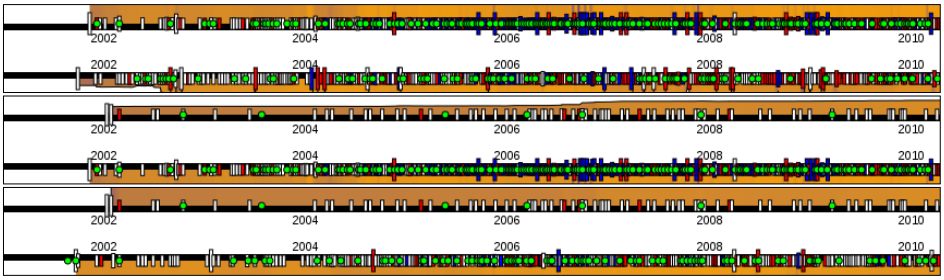


Figure 4.36: Timelines from the visualizations of articles on programming languages: C++ and BASIC (top), Emacs Lisp and C (middle) and Emacs Lisp and Lisp (bottom).

Emacs and Vi: In the UNIX world, there are several text editors, including the `Emacs` family (which includes `GNU Emacs`) and the `vi` family (which includes `Vim`).

Analyzing the plots for both articles (figure 4.37), we found out both articles had an edit rate increase around 2004, featuring several reverts and deletions, and exhibiting frequent talk page activity, probably a result of the everlasting "editor war" among the users of UNIX-like systems.

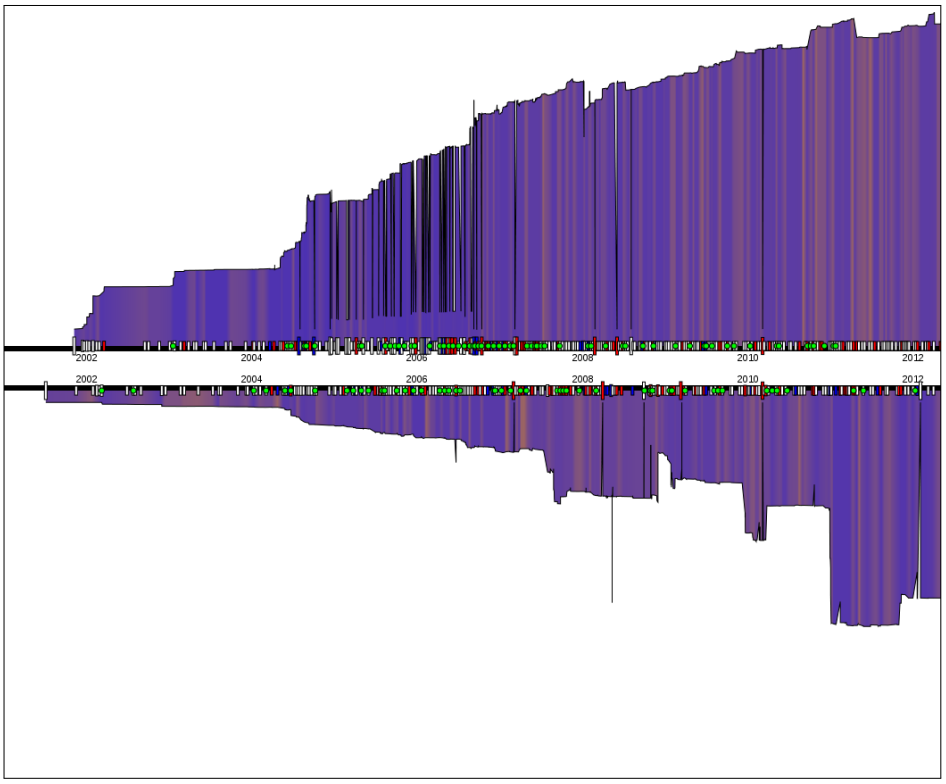


Figure 4.37: Visualization of the English Wikipedia articles on Text Editors: "Emacs" (above) and "vi" (below), set to plot the length, filled by controversy.

Interarticle vandalism: On 31 July, 2011, Stephen Colbert told the viewers of his show, "The Colbert Report", to "find the page on elephants on Wikipedia and create an entry that says the number of elephants has tripled in the last six months", after claiming to have changed Wikipedia himself to state that "Oregon is Idaho's Portugal"[13].

By inspecting the visualization of the articles on "Elephant" and "Portugal" between 20 July and 9 August, it became evident there was a sudden activity increase after Colbert's request, although activity on "Portugal" did not increase as much as it did on "Elephant" (figure 4.38). We were also able to find Wikipedia revisions 66987283[7], which replaced the content of the "Elephant" article with "The Number of elephants has TRIPLED in the past sixth months. – Stephen Colbert", and 67000659[8], which changed the introductory section of "Portugal" to read "Portugal, [...] is located between the Pacific ocean and the state of Idaho. Portugal is bordered by Washington to the north and by California to the south.".

[7] `http://en.wikipedia.org/w/index.php?oldid=66987283`
[8] `http://en.wikipedia.org/w/index.php?oldid=67000659`

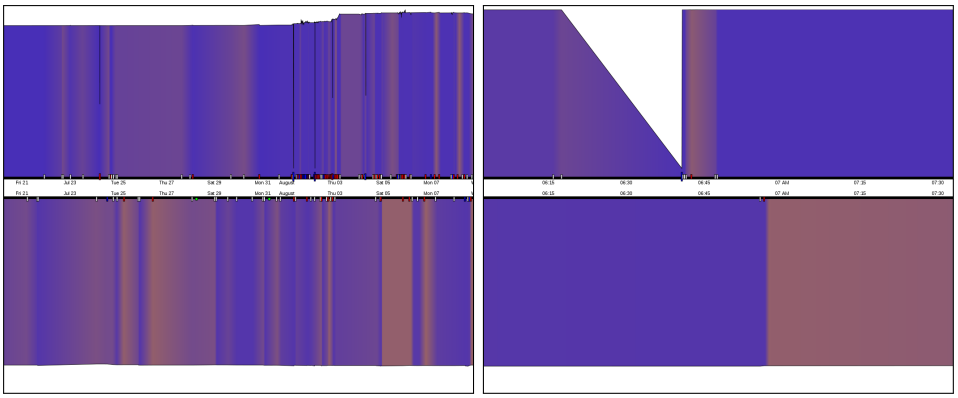


Figure 4.38: Visualization of the English Wikipedia articles on “Elephant” (above) and “Portugal” (below), with both plots set to plot length and fill by controversy: 20 July to 9 August 2006, showing an activity increase in “Elephant” (left), and detail of 1 August, where the revert in “Elephant” and deletion in “Portugal” correspond to the perpetrated vandalism (right).

4.4.3 Comparing Wikipedias

As our visualization allows the comparison of two independent articles from different MediaWiki instances, we have also explored the possibilities of interwiki comparison.

Similar activity: Depending on the subset of the regional community that is interested in the topic, it is possible that an article is updated at two different Wikipedias in a similar way. In the English and Hungarian articles on the former Hungarian president Schmitt Pál[9], depicted in figure 4.39, we found that significant activity changes spanned both Wikipedias: the sudden interest in the article at the end of June 2010, which also led to larger article sizes in both Wikipedias, is probably due to his nomination and subsequent election, on 29 June, as President of Hungary.

[9]Respectively, “Pál Schmitt” and “Schmitt Pál”, where the former has his name written in western name order.

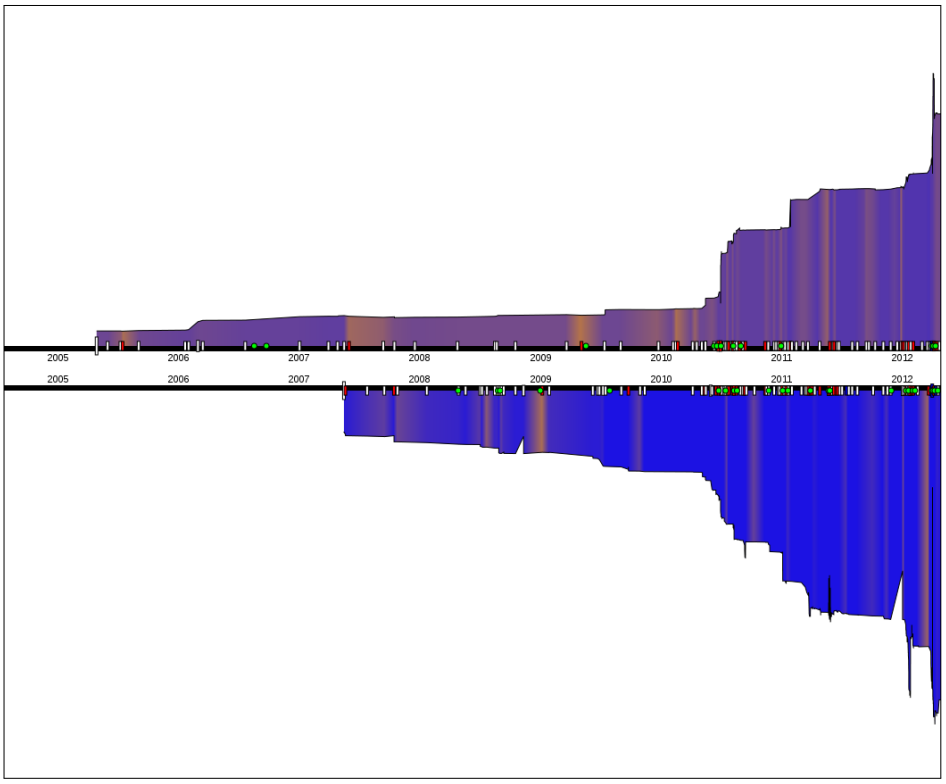


Figure 4.39: Visualization of the English Wikipedia (above) and Magyar Wikipedia (below) articles on “Schmitt Pál”, with both plots set to plot length and fill by controversy.

Taking a closer look at activities in 2012, two other bursts become evident: January and April. First, this shows a correlation between Wikipedia article activity and events outside Wikipedia: On 11 January 2012, a magazine accused Schmitt of plagiarizing his doctor thesis, explaining the first burst; On 27 March, the senate of his university was advised to withdraw his title, which it did on 29 March, and after which Schmitt announced his resignation from the office of President of Hungary, on 2 April, matching the second burst.

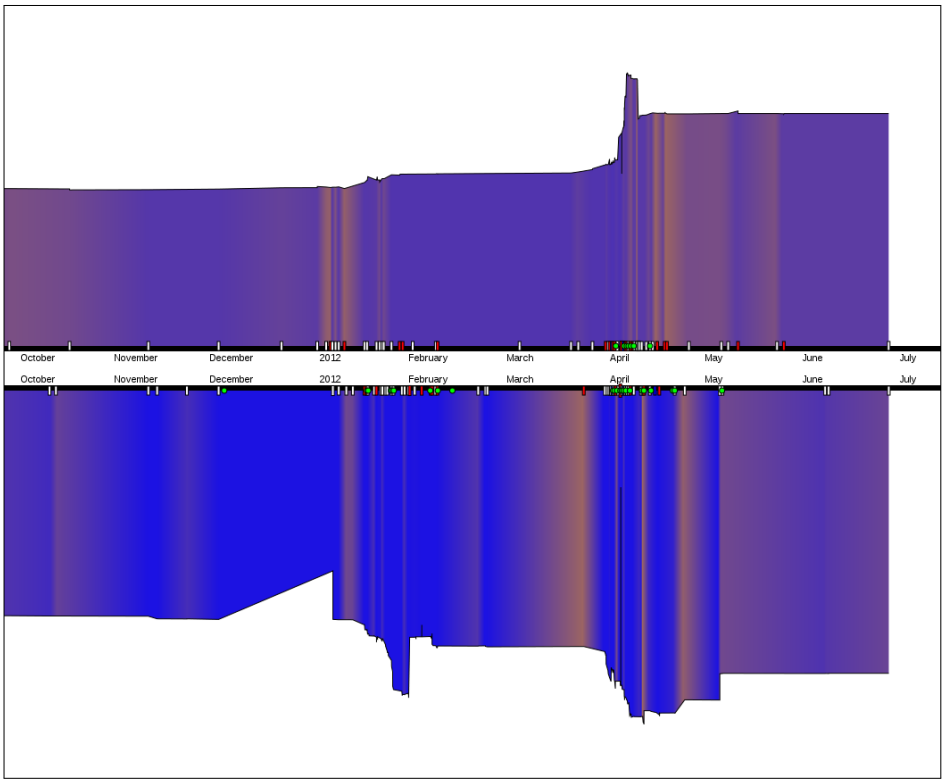


Figure 4.40: Visualization of figure 4.39, focusing on the last months of 2011 and on 2012.

Second, this again shows that the activity pattern is similar across both Wikipedias. Comparing the English article to the Portuguese one (with the same title), we get a different scenario. Unlike the English and Hungarian Wikipedias, the claims of plagiarism did not translate into heavy activity bursts, although some editors promptly updated the article. In figure 4.41, this comparison is shown, with the plot colored by author. As gray corresponds to the less frequent editors, we can conclude the Portuguese article does not have as many interested contributors as the Hungarian one.

As English is frequently used and taught internationally as a second language, it is possible some of the contributors who participate in the Hungarian article also collaborate in the English version. It is also possible that the English Wikipedia has more editors devoted to the coverage of current events in general.

Due to this international nature of the English language, the English Wikipedia may be seen as an international version of Wikipedia, rather than an edition constrained to a specific geographical area. Thus, we refrain from associating the higher activity with a popularity rise in English-speaking countries. On the other hand, it is possible for us to say that the topic appears to be much more popular among Hungarians[10] speakers than among Portuguese speakers.

[10]Hungarian is an official language only in Hungary, although it also has official status in Slovenia, Serbia, Austria, Croatia, Romania, Ukraine and Slovakia.

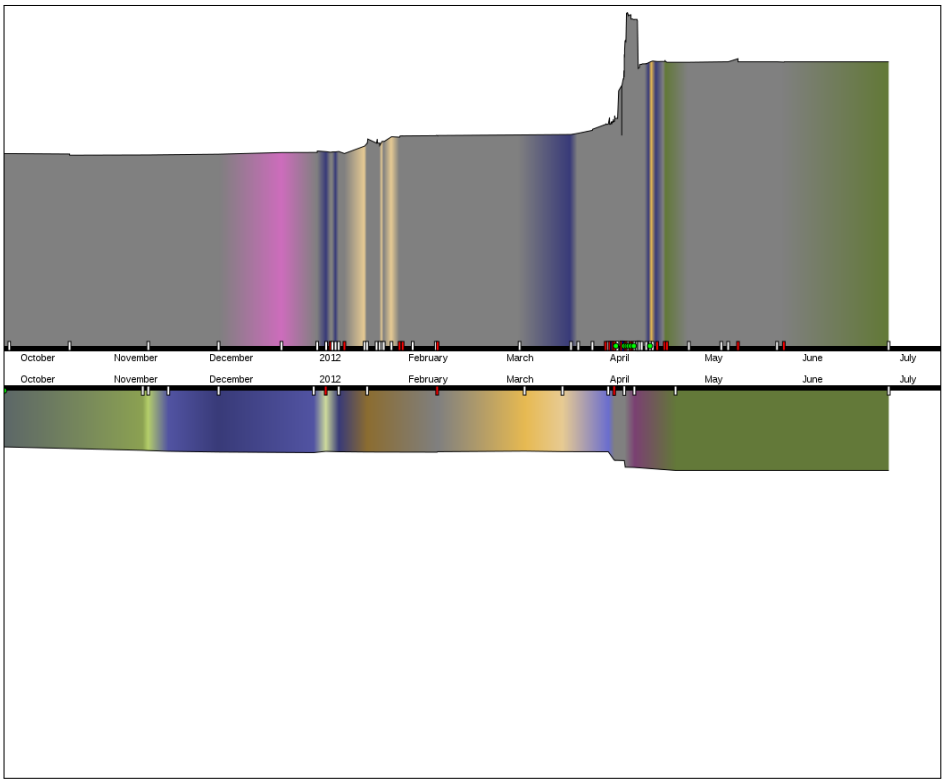


Figure 4.41: Visualization of the English Wikipedia (above) and Portuguese Wikipedia (below) articles on "Pál Schmitt", focusing on the last months of 2011 and on 2012, with both plots set to plot length and fill by author.

Different activity: On 10 April 2011, Fernando Nobre accepted an invitation to head the list of a political party for the electoral circle of Lisbon, with the explicit goal of being then elected president of the Assembly of the Republic[12]. The decision of the party leader to invite him, his acceptance and the bold way in which the leader of a party tried to choose the holder of the second most important political office in the country were heavily criticized and spurred controversy, ending with the deputies refusing to elect him[27].

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 (figure 4.42 left), which, if we inspect closely, did only occur in the Portuguese article (figure 4.42 right).

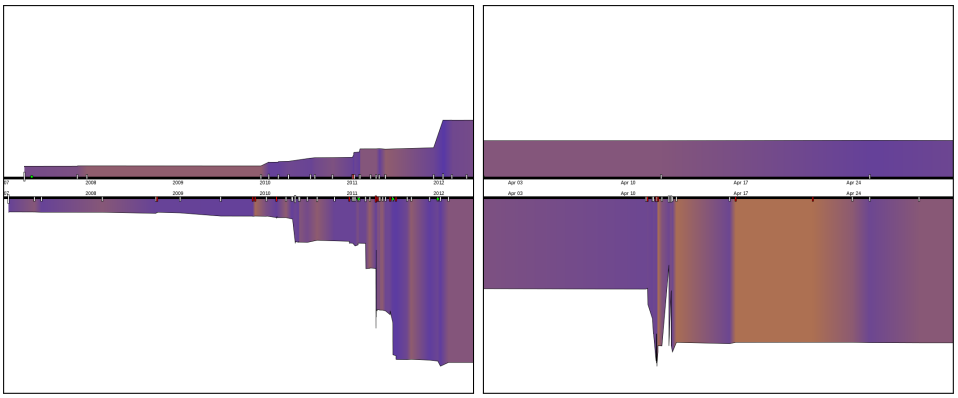


Figure 4.42: Visualization of the English Wikipedia (above) and Portuguese Wikipedia (below) articles on “Fernando Nobre”, first covering the entire lifespan of the articles (left) and then focusing on April 2011 (right). Both visualizations are set to plot length and fill by controversy.

Another example of a topic which exhibits different activity levels across different sites is Kalevala, a compilation of Finnish and Karelian epic poetry, whose article on the Portuguese Wikipedia is significantly less active than its counterpart from the Finnish Wikipedia (figure 4.43)

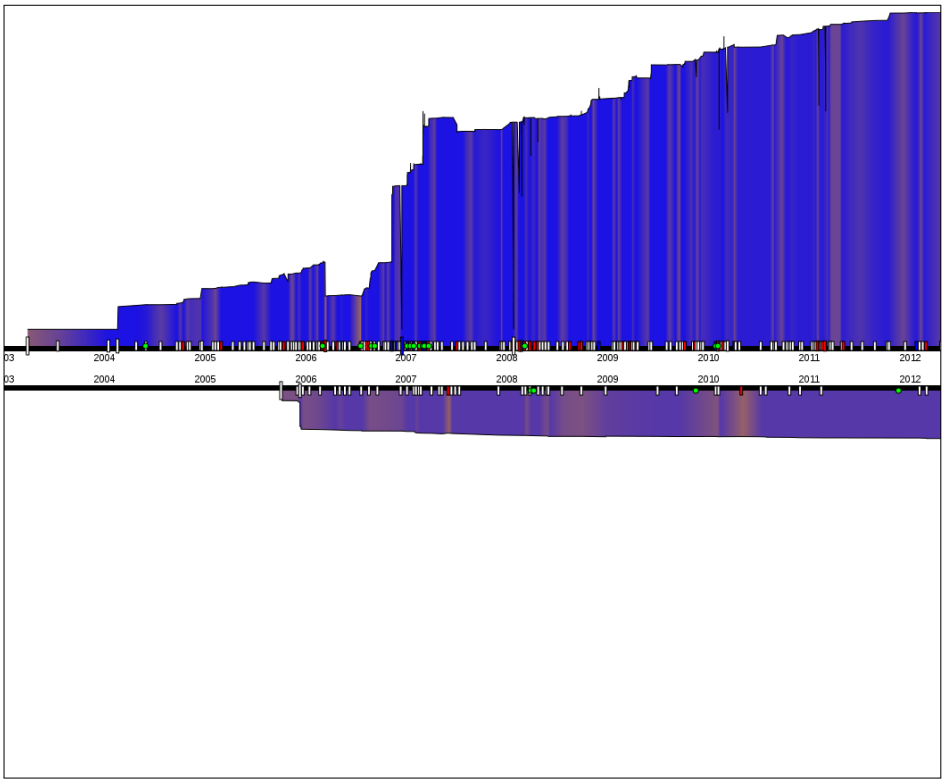


Figure 4.43: Visualization of the Finnish Wikipedia (above) and Portuguese Wikipedia (below) articles on “Kalevala”, set to plot length and fill by controversy.

The same distinction is found when comparing the Finnish and Portuguese articles on “Os Lusíadas”, an epic poem on the Portuguese voyages during the “Age of Discoveries”. The article from Portuguese Wikipedia is now more active than the corresponding article from the Finnish Wikipedia (figure 4.44).

In both cases, we believe it is safe to suggest that this reflects the popularity of those works in Finland[11] and in the Portuguese-speaking countries: "Kalevala" is known and popular in Finland, while "Os Lusíadas" is apparently virtually unknown there; the latter is frequently discussed in Portuguese countries (for example, in Portugal, where it is included in compulsory education curricula).

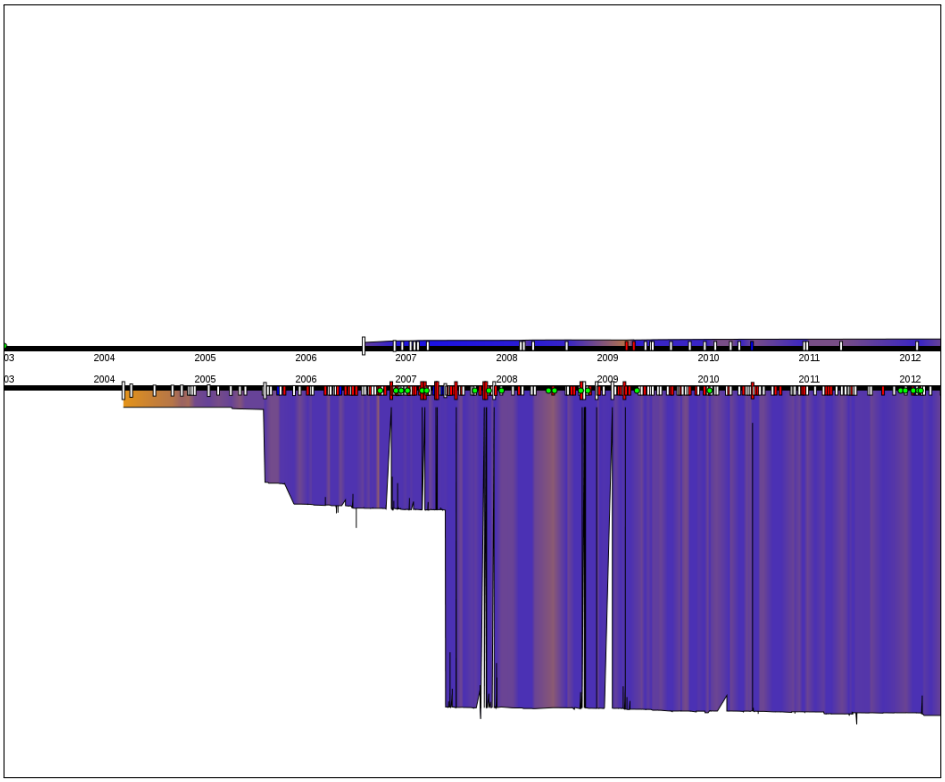


Figure 4.44: Visualization of the Finnish Wikipedia (above) and Portuguese Wikipedia (below) articles on "Os Lusíadas", set to plot length and fill by controversy.

[11] Finland is the only country where Finnish is an official language, although it is also officially considered a minority language in Sweden and Karelia (Russia).

Chapter 5

Evaluation

In this work, we presented a novel visualization to convey the revision history of a Wikipedia article, whose success we will now evaluate in the current chapter.

As the usefulness of the visualization part of our work relies on the interpretation by the users of the visualization screen, our main concern when evaluating the visualization is finding out whether users are able to find known patterns through the interface, and whether users can unveil previously unknown patterns.

As the main contribution of our work lies on the visualization, we chose not to perform any evaluation on the metrics component, focusing only on the evaluation of the visualization which presents those metrics to the user.

In order to assess the suitability of the visualization, we have to carry out user tests, recording whether users are able to spot a specific set of patterns and events, and also recording additional events users detect on their own.

In some of the surveyed visualization works, authors conducted evaluations of the resulting system or algorithm. Looking at the methods used in these works (WikipediaViz, iChase, ThemeRiver and Omnipedia), we found that all authors had chosen a similar approach, evaluating their own system through user tests (whose details are listed in table 5.1).

When evaluating WikipediaViz, as its authors employed hypothesis testing in the analysis of the test results, users were not asked open questions, they were instead asked to rank pages and to fill a questionnaire before participating in the test. WikipediaViz also had more test subjects than both iChase and ThemeRiver, possibly in an attempt to ensure the statistical significance of the conclusions of their analysis.

Inspired on these approaches, we decided to extend our user test protocol with questionnaires and a preparation step, where users are introduced to some of the visualization system concepts.

	WikipediaViz[11]	iChase[26]	ThemeRiver[19]	Omnipedia[6]
Number of subjects	24	8	2	27
Subject role	Not contributors	8 contributors (4 administrators)	—	Readers
Introduction?	✓	✓	✓	✓
Preliminary questions?	✓			
Activity	Ranking	Open questions	Open questions	Open questions, metrics
Statistical analysis?	✓			✓

Table 5.1: Details on the user tests carried by some authors. (As ThemeRiver is not centered on Wikipedia, there is no information about the role of the study subjects.)

5.1 Methodology

The core of the user tests consists on a set of scenarios, which comprise several tasks, concerning the usage of our visualization to spot known patterns or events in known controversial articles, and to relate patterns along two articles.

These tasks were presented to the users after some introductory steps, which included a verbal explanation of the goals and of the motivation behind the visualization, as presented in chapter 1; an introduction to and demonstration of the visualization features and elements, which we listed in section 4.2; and some minutes of unmonitored experimentation time, where users were allowed to use the visualization and ask questions on its elements, in order to make sure they were already used to the most important features when executing the tasks.

During the execution of the tasks, information was collected regarding the duration of the task, the number of mouse button clicks and keyboard key presses and the final outcome of each task (whether the users were successful and whether they committed any error).

Additional information was then collected through a written questionnaire, split in three parts:

- *personal questions* asked in order to assess the demographics of the inquired population, comprising gender, age, education level, computer experience and Wikipedia experience;
- *scaled questions on the system usability* asked using the System Usability Scale introduced by John Brooke[10], a generic Likert scale that allows a numeric evaluation of the system usability and its comparison with other systems evaluated using the same scale;
- *open questions* regarding the visualization usefulness and the fitness of the provided metrics, and any other comments and suggestions the inquired users want to share on the visualization.

This protocol was then tested and refined by doing two preliminary user tests, whose results were not included in the analysis presented in this chapter.

5.1.1 Scenarios

User tests comprised five scenarios, made of several tasks, which we describe in detail below, along with their purpose and success metrics.

Scenario 1: This scenario features tasks that were built around the English Wikipedia article "List of common misconceptions". As discussed in section 4.4.1, references in several popular media[23, 15] during early 2011 led to a sudden increase in the edit rate and article size.

Users were asked to:

1. Visualize the English Wikipedia article "List of Common Misconceptions";
2. Identify article activity changes and when did those happen;
3. Identify the contributors whose contributions survive longer.

With the second task, we wanted to find out whether users could identify the activity peak that was caused by the references to the article during January 2011. Although users may possibly try to use our metrics, as we depict the occurrence of changes over the timeline, this task is mostly intended to test whether users are able to understand the timeline, that is, identify parts of it which present an higher density of markers.

As the article was already active before January 2011, in a big picture view of the entire history, the January 2011 peak may be difficult to spot. Thus, this task will also show us if the visualization allows users to clearly tell apart activity spikes inside an already active period of an article.

In the third task, we ask users to identify the authors whose contributions survive for longer periods of time, thus assessing whether the visualization enables users to assess authorship information, through its "color: by author" plot area fill setting, ideally identifying the most active authors by their non-gray colors, and then assessing their names using the tooltip.

We found the two most frequently listed authors to be "Hippo43" and "Rracecarr", article-wise, and "ClueBot" and "SmackBot" section-wise.

Scenario 2: Involves an article comparison task, and the extraction of information from the content. It is based on the English Wikipedia articles "42" and "73", articles about the years 42 and 73 of the Julian calendar.

Even if not directly related, we found the article 42 to be a frequent target of vandalism and other controversial changes, regarding the popularity of the number. It plays a central role in Douglas Adams' novel "The Hitchhiker's Guide to the Galaxy", where it is presented as "The answer to the Ultimate Question of Life, the Universe, and Everything"[3].

Users were asked to fulfill the following tasks:

1. Visualize the English Wikipedia article "42";
2. Visualize the English Wikipedia article "73";
3. Rate the activity level of the article "42", when compared with "73" (faint, normal, excessive);
4. Identify an explanation for the activity level rating you chose.

With the third task, we want to find out whether the visualization enables users to successfully compare articles, by checking whether users consider that the article "42" is more active than the article "73", by comparing the density of markers in the timeline along time.

The fourth task, is concerned with the assessment of a justification (using the content pane) for the observed difference, in the case the user reports a difference. As we know, by inspection, that the increased activity was due to Adams' book, and thus our success metric will be whether users identify this relationship.

As extracting such a specific, detailed justification out of the entire history of the article may not be an easy task, we will also check whether users can detect vandalism and controversy as less detailed reasons for the higher activity rate.

With this, we will be able to tell whether the visualization succeeds at enabling users to extract information from content (considering all answers that involve vandalism and controversy, not just Adams' book), and to which extent users are able to dig deeper and find the connection with Adams' book.

Scenario 3: The third scenario concerns the identification of activity rate changes and the extraction of information from the periods when the activity rate changed, along with authorship information. That is, information extraction at a deeper level than in the two previous scenarios, where authorship and information were assessed for the article as a whole.

The tasks focus on the English Wikipedia article "Elephant", which, as discussed in section 4.4.2, was vandalized on 31 July 2006, following a request made by Stephen Colbert on his television show "The Colbert Report"[13].

We asked users to:

1. Visualize the English Wikipedia article "Elephant";
2. Look for abnormal activity regions;
3. Identify an explanation for the abnormal activity;
4. Identify the contributor who triggered the abnormal activity.

Our goal was to check whether users are able to find the revision 66989865[1] of the article on "Elephant" (in the second task, by first finding the higher timeline marker density, then zooming in and finding the sudden decrease in article length resulting from that revision), associating it with Stephen Colbert (third task, using the content pane in order to inspect the content) and attributing it to user "EvilBrak", the author of this revision (fourth task, using the tooltip).

Scenario 4: Then, we tested if users were able to spot activity changes when comparing two articles on the same topic from different Wikipedias: the English Wikipedia and Hungarian Wikipedia articles on Schmitt Pál ("Pál Schmitt" and "Schmitt Pál", respectively) these articles, previously shown as examples in section 4.4.3, were found to exhibit two peaks, on January 2012 and on March/April

[1] `http://en.wikipedia.org/w/index.php?oldid=66989865`

2012, which occur on both articles, starting first on the Hungarian Wikipedia, and are both possibly related with the accusations of academic misconduct directed at Schmitt.

Tasks were as follows:

1. Visualize, in one of the plots, the English Wikipedia article "Pál Schmitt";
2. Visualize, in the other plot, the Hungarian ("Magyar") Wikipedia article "Schmitt Pál";
3. Look for abnormal activity regions;
4. Of the regions you found, identify those that only occur in one of the Wikipedias;
5. For the regions which occur on both Wikipedias, identify the one where the abnormal activity started first.

With the third task, we once again intend to test whether the visualization enables users to spot changes in the activity rate, by pointing at least one of the 2012 peaks.

On the fourth task, we ask users if these peaks span both Wikipedias, thus assessing whether users are able to compare articles using the visualization.

Finally, we take article comparison to a further level: we ask users to point on which Wikipedia did those peaks start first.

All of these tasks involve the identification of zones with higher marker density in the timeline and also comparing those zones across the two timelines.

Scenario 5: As a last test, we asked users to compare and extract information from the Portuguese Wikipedia articles on the Portuguese comedians "Zé Diogo Quintela" and "Ricardo de Araújo Pereira", asking users to:

1. Visualize, in one of the plots, the Portuguese Wikipedia article "Ricardo de Araújo Pereira";
2. Visualize, in the other plot, the Portuguese Wikipedia article "Zé Diogo Quintela";
3. Check if there is any correlation between the activity rate of the two articles and, if so, identify a possible reason.

With the first task, we once again test whether users can compare two articles and find relations among their activity rates, by inspecting the timeline marker density and the evolution of article length along time. In this case, we found both articles to be related.

Then, we ask users to extract information (which would be done using the content pane), in another attempt to assess how well does our visualization help users with extracting information from the article content. In this case, we found the relation to be due to the participation of the two comedians in the comedy group "Gato Fedorento".

As this group and its members are part of the Portuguese popular culture, special care was taken when monitoring the execution of the scenario tasks, refusing reasons for the correlation that had not been extracted using the visualization.

5.2 User Profile

User tests were carried with a population of 20 users (16 male, 4 female), with an average age of 26 years (σ = 7.8, median 24 and mode 22), 8 of which were high school finalists, 9 bachelor degree holders and 3 had a master degree.

Users were asked to rate their own computer skills (1–5), having rated themselves as follows: most users (13) considered themselves as excellently skilled with computers (5); 5 users ranked their experience as 4; one user as 3 and one user as 2.

We also asked users to rate their Wikipedia skills (1–5): three users reported excellent skills (5), nine users rated themselves as 4, five users as 3, two users as 2 and one user as 1. Only one of the surveyed users reported having an account in at least one of the wiki sites operated by the Wikimedia Foundation.

Users' computer and Wikipedia skills are summarized in figure 5.1.

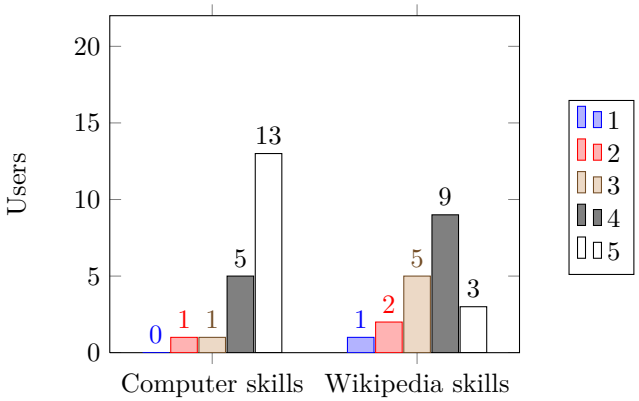


Figure 5.1: Summary of users' computer and Wikipedia skills.

5.3 Scenarios

All scenarios include one or more article visualization tasks. These are simple tasks, whose only purpose is to ask the user to load the articles the remaining scenario tasks rely on. As these visualization tasks are similar and we did not observe any issue with their execution, we will not discuss these tasks in detail below, focusing only on the remaining tasks.

In the visualization tasks, users were not found to have issues with the interface, quickly managing to pull the specified articles. Some users, probably due to a matter of personal taste, did not rely on the autocomplete suggestions, but the users who chose to rely on those were equally successful. We interpret this as an evidence that the autocomplete feature fulfills its goals.

5.3.1 Scenario 1

In this scenario, which took users, on average, 726.4 s to complete, users were asked to:

1. Visualize the English Wikipedia article "List of Common Misconceptions";
2. Identify article activity changes and when did these happen;
3. Identify the contributors whose contributions survive longer.

In the second task, meant to assess whether users would be able to identify the activity peak (timeline marker density peak) caused by the references to the article during January 2011, only 9 of the 20 users were successful, although 8 of the remaining 11 still identified other peaks. Only 3 users were unable to spot any difference in the activity rate.

This strongly suggests our visualization is able to fulfill the goal of enabling its users to identify differences in editor activity along time; on the other hand, the less clear result on the 2011 peak leads us to conclude that our visualization does not allow users to clearly tell apart activity spikes in already active articles, as is the case of this article (figure 5.2).

Figure 5.2: Timeline for the entire revision history for the English Wikipedia article "List of common misconceptions". Note how, despite the sudden increase in the activity rate that occurred on January 2011, that increase is not clearly assessable from the timeline.

Users were then asked, in the third task, to identify the authors whose contributions survive for longer periods of time. Only four users were able to identify at least one of the two most frequent authors (either article-wise or section-wise, those authors being "Hippo43", "Rracecarr", "ClueBot" and "SmackBot").

If we also consider "A Quest For Knowledge" or "Cresix", respectively the third and fourth most frequent authors, article-wise, which were identified by 10 users, 6 of which had not identified any of the four authors above.

We expected users to rely on the "color: by author" fill setting, and this result points out that our visualization, in general, and this fill setting, in particular, may be unsuitable for a big-picture analysis of authorship.

Users were found to do between 10 and 93 mouse clicks ($\overline{x} = 36.75$) and to press from 25 to 90 keys ($\overline{x} = 46.15$, with the name of the article being 29 characters long). Differences between minima and maxima are probably due to the way users interacted with the interface: some users entered the complete article name, others just wrote the first characters, relying on the autocompletion feature; some users heavily relied on the mouse zoom in/out feature, while others also used the date range text inputs, entering dates with the keyboard.

A summary of success measurements by task is shown in figure 5.3. In this figure, as in the ones which follow for scenarios 2–5, we depict the number of users which identified the pattern or answer

we were expecting, “Successful (expected results)”. As this does not include users who identified other, similarly valid patterns, we also present the number of users who were successful in the task, if we consider all the valid answers (including the ones we were expecting), “Successful (all results)”.

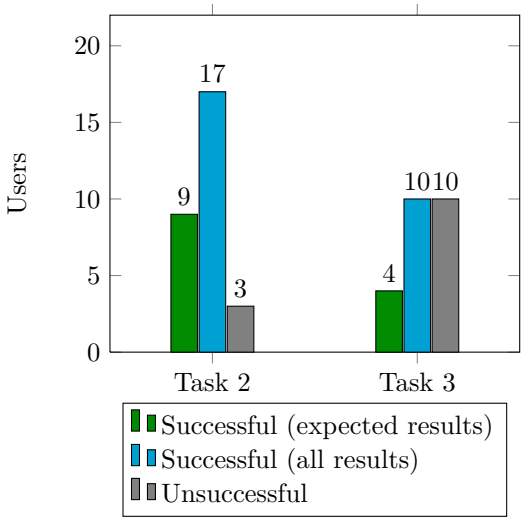


Figure 5.3: Summary of success measurements by task for Scenario 1.

5.3.2 Scenario 2

This scenario took, on average, 353.95 s to complete, and involved the following tasks:

1. Visualize the English Wikipedia article “42”;
2. Visualize the English Wikipedia article “73”;
3. Rate the activity level of the article “42”, when compared with “73” (faint, normal, excessive);
4. Identify an explanation for the activity level rating you chose.

The third task, concerning the comparison of the articles, was almost completely successful, with 19 of the 20 users considering that the article “42” was more active than the article “73” (the other user considered “73” to be more active), which proves that our visualization makes it easy for users to compare activity across different articles, by comparing the marker density of the timelines.

In the fourth task, concerned with the assessment of a justification for the observed difference, 10 users were able to justify it with vandalism or controversy (5) or with Douglas Adams’ book (5). Two users mentioned the higher number of editors and frequent updates to the list of references, one user said the article “42” was broader, and another linked the higher activity with the events that occurred in the Roman Empire (mentioned in the article section “Roman Empire”). This shows our visualization enables users to easily extract information from the content of older versions of articles, using the content pane.

Users were found to do between 7 and 86 mouse clicks ($\overline{x} = 33.65$) and to press from 5 to 44 keys ($\overline{x} = 15.00$, with the name of the articles totaling 4 characters). Again, we believe the broad range of values is due to how different users manipulated the visualization.

A summary of success measurements by task is shown in figure 5.4.

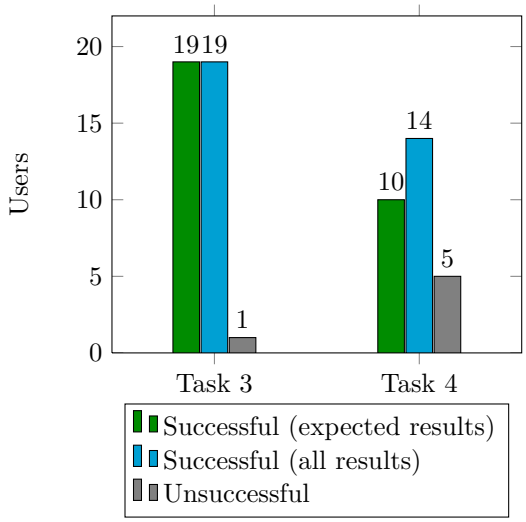


Figure 5.4: Summary of success measurements by task for Scenario 2.

5.3.3 Scenario 3

In this scenario, we asked users to:

1. Visualize the English Wikipedia article "Elephant";
2. Look for abnormal activity regions;
3. Identify an explanation for the abnormal activity;
4. Identify the contributor who triggered the abnormal activity.

Our goal was to check whether users are able to find the revision 66989865[2] of the article on "Elephant" (second task), associate it with Stephen Colbert (third task) and correctly identify its author, "EvilBrak" (fourth task).

With this goal in mind, the tasks were virtually unsuccessful, with only two users spotting the 2006 July/August peak. No user found the connection to Stephen Colbert or Wikipedia user "EvilBrak", although both users mentioned undesired changes (one attributed the increase to vandalism and the other mentioned was a sudden increase in the amount of deletions and reverts) as the reason for the unexpected activity.

On the other hand, only two users reported being unable to spot any activity change, with all other users identifying at least one time period (even if they did not spot the 2006 July/August one).

The remark we made for scenario 1 holds: this article is very active, thus possibly making it harder to spot abnormal activity from a high-level view of the article, as finding this specific edit would involve identifying the higher timeline marker density around that month, then zooming in the region and finally finding the sudden decrease in article length resulting from that revision. The high activity level of the article probably prevented users from spotting the change from an high-level view.

[2] `http://en.wikipedia.org/w/index.php?oldid=66989865`

The scenario took, on average, 685.15 s to complete, and users were found to do between 3 and 127 mouse clicks ($\overline{x} = 34.65$) and to press from 9 to 68 keys ($\overline{x} = 23.45$, with the name of the article totaling 8 characters).

Aside from the remarks already made for scenario 1 and 2 regarding these values, the higher maximum number of mouse clicks is possibly a consequence of the many times the chosen article was targeted by vandals. On the other hand, it should be noted that, despite the increase in the upper bound on the number of clicks, the average value is close to that of the two previous scenarios.

A summary of success measurements by task is shown in figure 5.5. We do not include tasks 3 and 4, as those depend on the identification, by users, of the specific revision we chose for task 2. Another consequence of this is that there are no successful users, if we only consider the specific revision we chose ("Successful (expected results)").

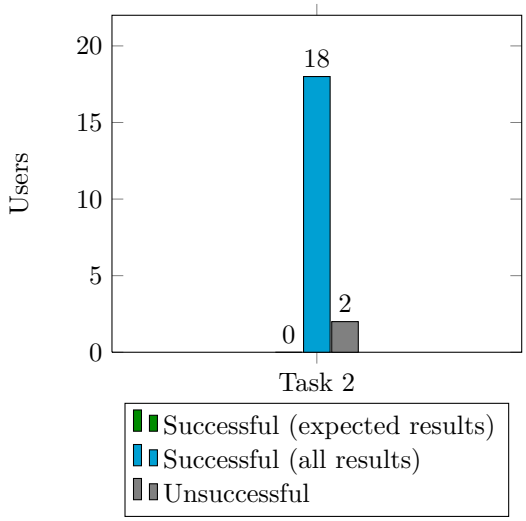


Figure 5.5: Summary of success measurements by task for Scenario 3.

5.3.4 Scenario 4

The fourth scenario, intended to test the assessment of activity rate changes and the comparison of different articles, involved the English Wikipedia and Hungarian Wikipedia articles on Schmitt Pál ("Pál Schmitt" and "Schmitt Pál", respectively).

Tasks were as follows:

1. Visualize, in one of the plots, the English Wikipedia article "Pál Schmitt";
2. Visualize, in the other plot, the Hungarian ("Magyar") Wikipedia article "Schmitt Pál";
3. Look for abnormal activity regions;
4. Of the regions you found, identify those that only occur in one of the Wikipedias;
5. For the regions which occur on both Wikipedias, identify the site where the abnormal activity started first.

Most users (14 of 20) were successful in identifying at least one of the activity peaks we detected: all of them identified the late March–early April 2012 peak, while 8 users spotted the January 2012 peak. All users were able to point out at least one time period.

We believe the not so high success rate of the third task for the January 2012 peak is due to this peak being not as bold as the March/April one: the latter translates in an article length spike and talk page activity, other than just an edit rate increase, while the January peak only shows an increased edit rate in one of the sites.

When asked whether those peaks span both Wikipedias, 5 users (in 8) incorrectly classified the first peak as having occurred in only one of the sites (4 said it only occurred in the Hungarian page, 1 said it only occurred in the English page); two users misclassified the second peak in the same way (1 English, 1 Hungarian).

Although we cannot point an exact reason for the remaining misclassifications, the first (January 2012) peak was only marked by a moderate edit activity increase in the English Wikipedia, while there was a greater activity increase, accompanied by talk page activity in the Hungarian Wikipedia, which explains the answers that considered the peak to have occurred only in the Hungarian page.

Finally, users were asked to point on which Wikipedia did those peaks start first. According to our own analysis, both peaks started earlier on the Hungarian Wikipedia, but 3 users answered otherwise, one of them for both peaks and the two other for the March peak. These fluctuations are more related to the way users interpret the timeline markers rather than deficiencies in the visualization: although for both peaks, those almost started at the same time, there are smaller differences in the edition timestamps, which some users chose to interpret as an indication those peaks did not start at the same time.

We expected users to execute these tasks relying on timeline markers and their density. Even considering the misclassifications and missed peaks, we believe the results clearly show that the timeline enables users to spot (some of) the peaks and to compare those across articles.

The scenario took, on average, 356.65 s to complete, and users were found to do between 3 and 77 mouse clicks ($\overline{x} = 24.10$) and to press from 15 to 86 keys ($\overline{x} = 39.25$, with the name of the articles totaling 22 characters (11 characters each)).

Once again, the average number of clicks is close to the averages observed in the previous scenarios. The lower average completion time, compared to scenarios 1 and 3, is probably due to the complexity of those scenarios: both involved information or metainformation retrieval, while this scenario was built solely around article comparison.

A summary of success measurements by task is shown in figure 5.6.

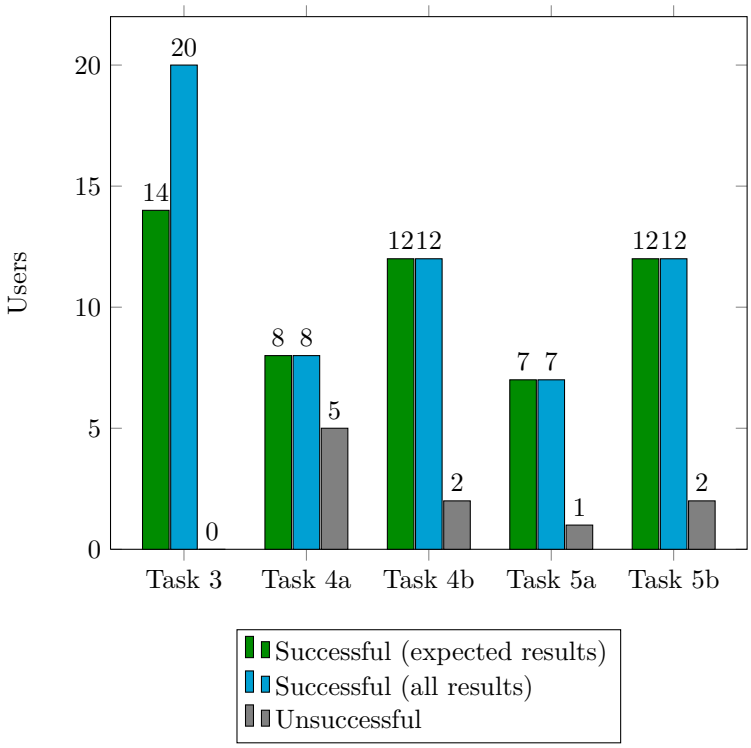


Figure 5.6: Summary of success measurements by task for Scenario 4. For tasks 4 and 5, a) refers to the success regarding the January 2012 peak and b) the March 2012 peak.

5.3.5 Scenario 5

In this scenario, we tested article comparison together with information extraction, asking users to perform the following tasks:

1. Visualize, in one of the plots, the Portuguese Wikipedia article "Ricardo de Araújo Pereira";
2. Visualize, in the other plot, the Portuguese Wikipedia article "Zé Diogo Quintela";
3. Check if there is any correlation between the activity rate of the two articles and, if so, identify a possible reason.

In the first task, except for 3 users, all users classified the articles as related, which is a good indication that our visualization fulfills the goal of enabling article comparison, both through the comparison of timeline marker density and of article length.

During the second task, 13 of the 17 users pointed out "Gato Fedorento" as the reason, one user found out their participation in "O Perfeito Anormal", another user just mentioned that the two comedians hosted a television show together and two other users found out the subjects work together. That is, all users were able to find a valid reason, once again showing that our system enables users to obtain information on the articles, through the content pane.

This scenario took, on average, 237.75 s to complete, and users were found to do between 8 and 58 mouse clicks ($\overline{x} = 23.05$) and to press from 26 to 110 keys ($\overline{x} = 50.35$, with the name of the articles totaling 44 characters).

It should be noted these values are closer to those of scenario 2, except for the number of key presses. The similarity is expected, as both scenarios involve similar tasks. The discrepancy in the number of key presses is mostly due to the length of the article names.

A summary of success measurements by task is shown in figure 5.7.

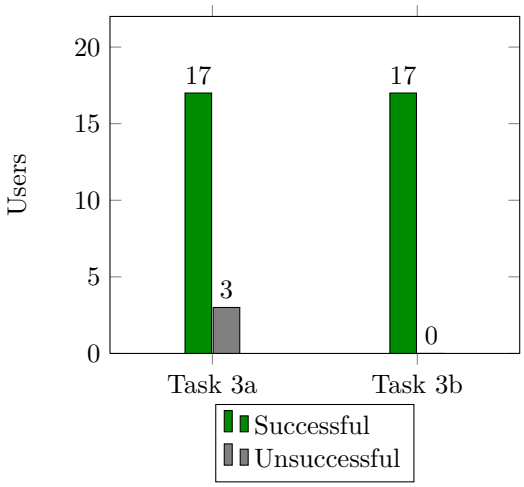


Figure 5.7: Summary of success measurements by task for Scenario 5. Task 3a refers to the identification of the relation between the articles, and Task 3b to the assessment of a justification for the relation.

5.4 User Reactions and Findings

All users considered the visualization fulfilled the goal of conveying information from a wikipedia article in a visual way, with one user noting that it will have to be used together with other methods if we want to confirm our findings.

Users were then asked which metrics did they find to be the most useful, with length being chosen by 11 users, followed by quality (7), controversy (6), vandalism (5) and color by author (4).

Metrics: Some users found metrics hard to understand, suggesting that we add some explanations or review the chosen metric names. As many scenarios involved increased activity, it was suggested that we include information on visit counts, and addition which would be, indeed, possible, using page view statistics made available by the Wikimedia Foundation[3].

One user pointed out that quality and controversy did not seem to be as reliable as the other metrics

Help screens: Several users suggested the addition of help screens or explanations, in order to help first-time users to understand the visualization elements. We believe this would improve the experience of first-time users, and should be considered if the system is ever deployed with the goal of serving infrequent users. A possibility could be the addition of small help buttons with associated help tooltips, although care must be taken to ensure such screens and explanations are unobtrusive, otherwise the visualization will be doomed to failure next to anyone but infrequent users. A user also

[3]http://dumps.wikimedia.org/other/pagecounts-raw/

suggested we add information on what action is currently associated to each of the mouse buttons and operations.

Timeline: Two users suggested the addition of a zoom-fit feature, that would change the visualization date range to encompass the entire article data. Although this is not possible for zooming out, as the visualization only receives data in the specified timerange, it is possible when zooming in. It was also suggested that a tooltip is shown when hovering a timeline marker. Although, with the new optimization step, one marker can comprise several edits, this would still be an interesting way to allow users to get more information directly from the timeline.

Usability: Some problems were pointed out, such as the detailed sensitivity of the tooltip locking feature, which, when the pointer is moved before releasing the mouse button, even if just for a couple pixels, triggers zoom instead and the lack of a way to cancel a pending request.

Users also complained about the unability of the AJAX search box to show the entire text of the results, as those were limited to the width of the selection list widget. This problem was present in the version of the visualization used for user tests, but has been fixed in the final version.

The lack of a way to relate the vertical axis scale from both plots was also pointed out as a flaw. Users suggested adding a background grid or explicitly showing the vertical axis with its range. In the final version, we introduced a feature that normalizes the vertical axis across the two plots, which makes values comparable between plots.

Hiding one of the plots when it is not used was suggested as a feature. Although not essential, this would help users focus on the plot they are analyzing.

A user also suggested that the date input fields should be complemented with a calendar-like mouse-based date picker.

Zoom itself was also pointed as an operation that could be improved, as it only worked over the plot area, but not on the white background. This was also fixed in the final version.

Aesthetics and appearance: Comments were also raised on whether it would be possible to find a better alignment between the plot options and the plots. Users complained the tooltip would be sometimes rendered unreadable, as there is no code in place to avoid getting its content cut at the bottom and right ends of the screen. One user found the gradient colors confusing and unintuitive.

Date under the pointer: During one of the tests, the idea of adding, in the tooltip, the exact date under the pointer was suggested. Although, in most cases, this date would not correspond to any edition, this would make it easier for users to extract dates from the visualization.

Simplification and integration of datapoints: A user pointed out that the system becomes harder to interact with and use when it tries to convey large amounts of information. Some optimizations, introduced after user tests, fix some of the speed and visualization issues caused by the huge volumes of information associated with some articles, although this does not exclude the possibility of

applying additional optimizations, in order to improve the visualization even more. It was also suggested that the visualization should convey derivatives of the metrics, or provide some derivative-based metric.

Color by author: Some users were expecting the color "by author" gradient to follow some key. Although the main idea was just to highlight the most frequent authors, we welcome the suggestion, which could translate at least on a key where each frequency is always assigned the same color, and possibly in a key which explicitly lists the authors currently associated with each color.

5.4.1 User findings

When analyzing the "Elephant" article, most users identified two length spikes, which correspond to two occurrences of vandalism:

On 8 December 2005, an anonymous user (connecting through a computer identified by the Internet Protocol address `138.87.135.25`) added numerous repetitions of the expression "ELEPHANTS R COOLOOOOOOOOOOOHHHHH[…]HHH" to the article (revision 30603285[4]);

On 10 July 2007, user Kasdun added a "shock image" to the beginning of the article (revision 143754729[5]).

These two changes were easily spottable due to the drastic length increase, thus standing out when the graph is set to plot the article length over time.

Users also spotted other changes which, like the one we were expecting users to find, result in a sudden decrease, for a very short time, of the article length:

One user spotted revision 98352289[6], which is a late reaction to Colbert's request — Daniel James Jacob Nelson changed the article to read "The population of the African elephant has tripled over the past six months.", thus quoting Colbert.

During scenario 4 ("Pál Schmitt"/"Schmitt Pál"), 9 users also reported the change in mid-2010 we mention in section 4.4.3, possibly a result of Schmitt's election to the office of President of Hungary.

A user explored the article on the Blizzard Entertainment game "Diablo III" (figure 5.8), being able to spot two acts of vandalism: revisions 363563326[7] and 73144073[8]. The user also commented, when seeing the sudden activity increase on June 2008, that it was probably due to the official announcement of the project by Blizzard, which, until then, had not acknowledged the development of a new game in the "Diablo" series.

[4] `http://en.wikipedia.org/w/index.php?oldid=30603285`
[5] `http://en.wikipedia.org/w/index.php?oldid=143754729`
[6] `http://en.wikipedia.org/w/index.php?oldid=98352289`
[7] `http://en.wikipedia.org/w/index.php?oldid=363563326`
[8] `http://en.wikipedia.org/w/index.php?oldid=73144073`

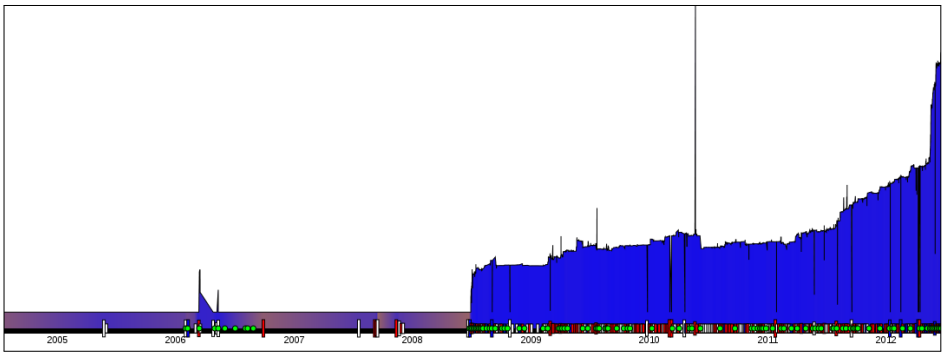


Figure 5.8: Visualization of the English Wikipedia article "Diablo III", where the plot is set to depict length, and is filled by quality.

Another user asked to explore the Portuguese Wikipedia article on "Foros de Salvaterra", a town in the Portuguese municipality of Salvaterra de Magos, as he already knew the article had been vandalized in the past, and wanted to see if he could spot that act of vandalism using our tool. His attempt was successful, as he found the change introduced in revision 23177781[9], where the anonymous user behind the IP address `89.152.23.16` added a paragraph stating this town is "two years ahead in the future", having thus already experienced the "end of the world" phenomenon some people mistakenly believe to be predicted by the Mayan calendar.

Although not mentioned by the user, it should be noted that this change was quickly undone by user ChristianH just one minute later, highlighting how quickly vandalism gets fixed even in less popular articles.

5.5 Questionnaire

Computing the final scores from the answers to the System Usability Scale questionnaires, following the instructions laid down by Brooke[10], scores were found to be, on average, 68 ($\sigma = 19$), median of 70 and mode of 60.

Analyzing the separate scores for each statement, summarized in table 5.2, and considering that SUS was designed by alternating positive statements (odd-numbered) with negative statements (even-numbered), we found out that all negative statements were, on average, disagreed with, even if some users "strongly agreed" with some of these statements. Positive statements were also mostly agreed with.

The not so high average score is probably due to the relative complexity of the visualization, which, while conveying information, does not do so in a simple way that appeals to infrequent users. Such values may, in general, be due to the fact none of the interviewed users carries out this type of analysis, and, in particular, to the possible inadequacy of some metrics or of their names.

[9] `http://pt.wikipedia.org/w/index.php?oldid=23177781`

	Statement	$\overline{x}$	$x \in$	median	mode	σ
1	I think that I would like to use this system frequently	3.2	1–5	3	3	1
2	I found the system unnecessarily complex	2.3	1–4	2	1	1.1
3	I thought the system was easy to use	3.5	1–4	4	4	0.8
4	I think that I would need the support of a technical person to be able to use this system	2.2	1–5	2	1	1.2
5	I found the various functions in this system were well integrated	4.2	3–5	4	4	0.7
6	I thought there was too much inconsistency in this system	1.6	1–3	1	1	0.7
7	I would imagine that most people would learn to use this system very quickly	3	1–5	3	2	1.2
8	I found the system very cumbersome to use	2	1–5	2	2	1
9	I felt very confident using the system	3.4	1–5	3.5	3	1.1
10	I needed to learn a lot of things before I could get going with this system	2	1–4	2	2	0.9

Table 5.2: Summary of the answers to the System Usability Scale questionnaire.

5.6 Discussion

Tests have shown us that our visualization fulfills its goals concerning the identification of activity patterns and comparison of articles. Information extraction was as successful as we would expect it to be, given the great volume of information available through the visualization.

We also saw that more active articles make it harder for users to spot activity changes. As, for several articles, there is much more information to convey than screen pixels to display it, not only did we opt for a top-down approach in the end, but we also discourage anyone attempting to follow a bottom-up approach from doing so, as it will not only cause interpretability issues, regarding hidden data, but will also lead to time and space complexity issues.

The devised tasks can be grouped in four major areas:

- *Spotting activity rate changes*, a goal covered by three tasks: scenario 1 task 2 (where 17 out of 20 users were successful), scenario 3 task 2 (18 of 20) and scenario 4 task 3 (20 of 20). That is, users were mostly successful, with an average success rate of, approximately, 91.67%;
- *Assessing authorship*, where only half of the users were able to correctly obtain information on the most frequent authors of an article (scenario 1 task 3, 10 users out of 20, success rate: 50%);
- *Comparing articles*, comprising four tasks (scenario 2 task 3 (where 19 users were successful, in 20), scenario 4 task 4 (part 1, with 8 successful users in 13; part 2 with 12 in 14), scenario 4 task 5 (part 1, 7 in 8; part 2 with 12 in 14) and scenario 5 task 3 (part 1: 17 on 20)), where users were, again, successful, with an average success rate of 83%.
- *Information extraction*, which encompasses two tasks (scenario 2 task 4 (14 users out of 19), scenario 5 task 3 (part 2: 17 of 17)) where users were asked to retrieve justifications or explanations for the observed patterns. Once again, users were successful, leading to an average success rate of 86%.

Tests have, thus, shown that the proposed visualization fulfills its goals concerning the identification of activity patterns, comparison of articles and information extraction, tasks which involve, mainly, the timeline and its markers and the content pane, also involving the analysis and comparison of the numeric metrics.

On the other hand, tests also showed that users had problems assessing authorship information (plot areas filled with "color: by author") This may be due to the the great volume of information available through the visualization, which hampers the fulfillment of this task, but is also evidence that this visualization feature needs to be redesigned, as its current approach is clearly inefficient.

The time-series approach we chose to guide the design of our visualization turned out to work as expected, effectively condensing an enormous amount of sequential data in a single screen.

Even if the time-series approach works and the plot line and area are able to convey information, the optimizations arising from a top-down simplified rendering approach would still improve the ability of the visualization to convey information. We believe the merging of datapoints and the use of a derivative-based approach should be investigated in future attempts to convey information in the same way.

It should also be noted that, although in the first stages, we chose a real-time approach for the computation of metrics, we soon realized that computations would be too time intensive and that we had underestimated both the cost of the retrieval of content from Wikipedia and the volume of data presented in a single visualization screen. Thus, we believe any work of this kind must rely on a cache-based approach, in order to avoid overloading the Wikimedia Foundation servers and to speed up our system.

Chapter 6

Conclusions

Driven by the belief that the growth of Wikipedia, given its openness (anyone can edit (most of) the articles), 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 and behaviors in an innovative way.

In order to accomplish this goal, related works were analyzed to assess the metrics used and the visualization techniques employed, which we then discussed in order to design our own solution.

Based on the insights from related works, we devised a set of metrics we then compute on a revision-by-revision basis, after retrieving the article content from Wikipedia.

Metrics are computed from the revision content, and are then cached in an on-disk database. This database is used to serve future requests for the same data. Along with metrics, we also store the retrieved revision content, in order to enable the real-time display of said content in the visualization.

The set of metrics is mostly based on linguistic features of the revision content, including compressibility, word frequencies and character frequencies. As word frequencies depend on the site language, although our work was prepared to work with different editions of Wikipedia, only a few sites will feature the complete version of these metrics, the ones for which we obtained lists of good, offensive and biased words.

We designed the entire system in a site-agnostic way: other than the site language, for which only a small subset of the Wikimedia Foundation-operated sites are recognized, there is no tie to a specific site or language. The entire data flow is built upon a small set of parameters that identify the MediaWiki server: its hostname and the path, in the HTTP server, to the MediaWiki instance.

Although our work focuses on the analysis of Wikipedia articles, our system can effectively be used to analyze articles from other sites, such as Uncyclopedia[1], the official Gentoo Wiki[2] or any of the Wikimedia Foundation "Sister Projects", such as Wikibooks[3].

These metrics, which comprise a time series, are then used to build a visualization focused on depicting the evolution of these metrics through time, while also enabling quick access to revision content.

[1] `http://uncyclopedia.wikia.com/`
[2] `http://wiki.gentoo.org/`
[3] `http://en.wikibooks.org/`

Our visualization comprises an AJAX-based web page, where plots are drawn using the `d3.js` graphics library, using data retrieved from another component, whose purpose is to compute and cache the article metrics.

For each revision, the visualization receives the metric values, and uses those to draw the plot line and define the color fill of the plot area. In some cases, the data may be condensed in order to avoid superposed graphical elements.

Plots are accompanied by their timelines, where article and talk page changes are marked along time, thus providing a quick visual depiction of the article activity along time.

We then analyzed several case studies, where we could find patterns, relate changes to real-world events and compare articles on different topics.

We carried user tests, finding out 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. It was also observed that users could successfully extract information through the visualization.

The tests also highlighted the need for some graphical optimization, in order to handle hidden data and resolve time and space complexity issues in the rendering stage.

Test results show that the time-series approach we chose to guide the design of our visualization turned out to work as expected, condensing a possibly large volume of sequential data in a single screen.

6.1 Future Work

During user tests, we collected several suggestions from users, some of which remain as future work, such as user interface improvements involving help screens, calendar-based date pickers and a zoom-fit feature.

As we monitored user tests, we realized that the system would need to be optimized in order to avoid performance and scalability issues. Some optimizations have already been implemented, but a proper analysis and assessment of possible additional improvements would be the next logical step.

Although our system is able to work with any API-enabled MediaWiki-powered wiki site, we did not explore sites other than some of the Wikipedia editions operated by the Wikimedia Foundation. Considering the increasing popularity of the wiki approach, we believe the application of this visualization to other sites and the study of patterns found in their articles would make for a promising research topic by itself.

We observed that some users had trouble understanding and interpreting our metrics. Despite being based on metrics from related works, our work aggregates these metrics from other works through a plain arithmetic mean, simply based on expectations regarding quality content, acts of vandalism and controversial editions.

A possible improvement would be the exhaustive analysis of the adequateness of the averaging and of the formulas, possibly leading to weighted means or even new, corrected formulas. This improvement could possibly involve the automatic assessment of an optimal set of weights through machine learning.

Yet another suggestion from users was the introduction of derivative-based information. While assessing the adequateness of metrics, tests could also be carried to test several derivative-based approaches, and other similar techniques which try to aggregate information from more than one revision when computing the metric for a single revision. Derivatives could also be employed in the visualization itself, researching the depiction of the variation of metrics instead of metrics themselves.

It is also possible that completely novel metrics could be devised for this system. As our visualization component merely retrieves information from the metrics assessment component, we would just need to make minor changes to the visualization component in order to include more metrics, while the major changes would have to be done just in the metrics assessment component.

The results of the user tests clearly show that our depiction of authorship does not perform well, not being scalable and failing to provide a quick and efficient way to assess authorship. Thus, this would be yet another direction for further research, possibly leading to a different authorship-based feature.

Currently, differences can only be shown across two consecutive revisions. Although that would be a costlier operation, removing this restriction, allowing the display of differences between any two revisions of the article, would lead to a more powerful visualization, where content analysis is concerned.

The major bottleneck regarding disk space complexity in our work is the disk cache, which can easily grow above several gigabytes for a small collection of articles, as the entire content for all revisions is stored, along with the differences between consecutive revisions. Finding a way to significantly reduce the required space would be a major improvement, as it would not only reduce the disk space requirements, but also enable the online deployment of our work without serious concerns regarding the available disk quota.

Appendix A

Wikipedia glossary

Wikipedia has pages, which can be read and edited by its **users**. When a user edits a page, she is said to be an *editor*, *contributor* or *author*. These terms are used interchangeably in this text. The works of **editors** are called **editions** (which is frequently shortened as *edits*) or *revisions*. These are sometimes referred to as *contributions* or *changes*.

These editions, which can be formally defined as a pair of consecutive versions of the same article[25], are kept by Wikipedia, even when there are newer editions, thus making the **edit history** (or *revision history*). In this history, editions may have an associated **comment**, which hopefully explains or describes the changes made.

Wikipedia pages are of many kinds. **Article pages** carry the main, encyclopedic content and have a corresponding **article talk page** where editions and other article-related subjects are discussed. There are also **user pages**, which contain information on the user, such as affiliation with Wikipedia subprojects, and have corresponding **user talk pages**, used for communication among users.

There are also other types of pages, mostly for administration and maintenance. Pages may carry banners or be protected when needed: **banners** display a visible, highlighted message to warn the user about some issue with the article; **protections** prevent sets of users from editing the page.

Pages can also be *tagged*, for example with a **dispute tag**, and be sorted into **categories**. **Templates** are pages carrying reusable blocks to be used in other pages (such as a train network map to be used in pages on the train stops). If an article is too short, it is said to be a **stub**, and is categorized and tagged as such.

Users, who can be either *anonymous* (identified only by the Internet Protocol address they use) or *registered* (identified by their name), and may be **administrators** or *robots* (shortened as **bots**), may get engaged in **conflicts** (also called *disputes*), of which **edit wars** are an example, characterized by consecutive **reverts** (attempts to undo some edition). Users may also get involved in **vandalism**, being then called **vandals**, hunted by **vandal fighters** and responsible for events such as **mass deletions**. Users can also be **blocked**.

Appendix B

Lists of Words

Here we present the lists of good, offensive and biased words for languages other than English:

- French Wikipedia
  - **"good words"**: 'de', 'la', 'le', 'et', 'en', 'du', 'des', 'les', 'est', 'dans', 'un', 'par', 'au', 'une', 'pour', 'il', 'sur', 'qui', 'que', 'a', 'son', 'avec', 'plus', 'se', 'sont', 'ou', 'ce' and 'aux';
  - **"offensive words"**[43]: 'beauf', 'blueneck', 'boche', 'con', 'enfoiré', 'fif', 'gabacho', 'gouine', 'métèque', 'nègre', 'polard', 'poufiasse', 'pédé', 'pétroleuse', 'racaille', 'social-traître', 'travelo' and 'wackes';
  - **"biased words"**[45]: 'souvent', 'généralement', 'inégalé', 'prétendre', 'soi-disant', 'naturellement', 'manifestement', 'essentiellement', 'principalement', 'extrême', 'terroriste', 'terrorisme', 'résistance', 'fondamentalisme', 'théorie', 'mythe', 'dictateur' and 'auteur'.
- Portuguese Wikipedia
  - **"good words"**: 'de', 'a', 'e', 'o', 'do', 'da', 'em', 'que', 'no', 'com', 'uma', 'um', 'para', 'na', 'os', 'por', 'dos', 'como', 'se', 'foi', 'as', 'mais', 'ao', 'sua', 'das', 'seu', 'ou' and 'ser';
  - **"offensive words"**[36]: 'babaca', 'bacanal', 'bacurinha', 'badalo', 'badamerda', 'baranga', 'bardamerda', 'barrote', 'bedamerda', 'benga', 'berdamerda', 'bichana', 'bilha', 'bimba', 'bimbada', 'boceta', 'boiola', 'bolagato', 'brichote', 'broche', 'bruaca', 'buceta', 'bugiranga', 'bunda', 'burro', 'cabaço', 'cacete', 'cachorra', 'cadela', 'cagalhão', 'cagar', 'canhão', 'capô-de-fusca', 'caralho', 'catatau', 'catraia', 'chavasca', 'chuchu', 'chupeta', 'colhão', 'comua', 'cona', 'conanas', 'conaça', 'coxanga', 'crica', 'cu', 'cu-de-ferro', 'cu-de-mãe-joana', 'cunete', 'cunilíngua', 'cuzão', 'fiote', 'foda', 'foder', 'forunfar', 'fuampa', 'fufa', 'furico', 'furunfar', 'gaita', 'gajo', 'galinha', 'grelo', 'grão', 'gulosa', 'idiota', 'jeba', 'jumento', 'kiwi', 'lola', 'mamado', 'mangalho', 'maricas', 'mastro', 'meinha', 'merda', 'messalina', 'minete', 'moca', 'nhonhoca', 'pandeiro', 'paneleiro', 'patolar', 'pau-no-cu', 'peida', 'peidar', 'peido', 'pene', 'pentelho', 'pica', 'pindocar', 'pinto', 'pintudo', 'piranha', 'piroca', 'pirrola', 'pito', 'piupiu', 'pixa', 'piça', 'porra', 'potranca', 'punheta', 'puta', 'puto', 'queca', 'quenga', 'rabeta', 'racha', 'rapidinha', 'sapatona', 'sapatão', 'siririca', 'suruba', 'tabaca', 'tanajura', 'tesudo', 'tesão', 'teta', 'toba', 'troca-troca', 'vadia', 'xana', 'xarifa', 'xarola', 'xarolo', 'xavasca', 'xereca', 'xibungo', 'ximbica', 'xiranha', 'xiri', 'xota', 'xoxota' and 'óbu';

- **"biased words"**[44]: 'lendário', 'grande', 'eminente', 'visionário', 'notável', 'líder', 'célebre', 'extraordinário', 'brilhante', 'famoso', 'renomado', 'prestigioso', 'respeitado', 'virtuoso', 'culto', 'racista', 'perverso', 'seita', 'fundamentalista', 'herege', 'extremista', 'negacionista', 'terrorista', 'libertador', 'controverso', 'acredita-se', 'suposto', 'alegado', 'pretenso', 'acusado', 'chamado', 'notavelmente', 'interessantemente', 'claramente', 'certamente', 'afortunadamente', 'felizmente', 'infelizmente', 'tragicamente', 'precocemente', 'revelou', 'indicou', 'expôs', 'explicou', 'encontrou', 'notou', 'observou', 'insistiu', 'especulou', 'conjeturou', 'alegou', 'afirmou', 'admitiu', 'confessou' and 'negou'.

- Spanish Wikipedia
  - **"good words"**: 'de', 'la', 'en', 'el', 'y', 'a', 'que', 'del', 'los', 'se', 'por', 'con', 'las', 'un', 'su', 'una', 'al', 'como', 'para', 'es', 'no', 'fue', 'o', 'lo', 'sus', 'entre', 'este' and 'esta';
  - **"offensive words"**[35]: 'alampar', 'batir', 'blanquillo', 'bola', 'bollo', 'cabronazo', 'cachada', 'cachar', 'cachero', 'cachucha', 'cacorro', 'cagar', 'cagarla', 'cagón', 'cajeta', 'callampa', 'calzar', 'caraja', 'carajo', 'cepillar', 'chele', 'chichi', 'chinga', 'chingadazo', 'chingadera', 'chingado', 'chingar', 'chocho', 'cholga', 'choro', 'chota', 'choto', 'chucha', 'chuchumeca', 'chupe', 'cimbrel', 'cipote', 'coco', 'cojón', 'concha', 'conchudo', 'correrse', 'coño', 'criatura', 'culero', 'culiado', 'culicagado', 'culo', 'cámara', 'difarear', 'difariar', 'diuca', 'empergeñar', 'empernar', 'enculado', 'escoñetar', 'fleto', 'follar', 'furcia', 'gilipollas', 'goma', 'gonorrea', 'güila', 'haiga', 'haigamos', 'haigan', 'haigas', 'haigáis', 'hijuna', 'hocico', 'hostia', 'hoyo', 'huevón', 'inflar', 'jalar', 'jarioso', 'joder', 'lagarta', 'leche', 'lumi', 'macaquero', 'machete', 'madrazo', 'madrear', 'madriza', 'maricón', 'marihuanero', 'mazo', 'mear', 'mes', 'mierda', 'mojar', 'mojarse', 'mojón', 'mostacero', 'nabo', 'nardo', 'orto', 'paja', 'paja', 'mental', 'pajero', 'pajilla', 'pajillero', 'panocha', 'pechuga', 'pedar', 'peder', 'pedo', 'pedorrear', 'perra', 'petar', 'picha', 'picho', 'pichula', 'pico', 'pijo', 'pinche', 'pindonga', 'pinga', 'pisar', 'playo', 'polla', 'polludo', 'poronga', 'potorro', 'pucha', 'puta', 'putazo', 'putañero', 'putiza', 'puto', 'pájaro', 'quilombo', 'raja', 'reculiado', 'sapo', 'tango', 'torta', 'tragasable', 'tragasables', 'tripa', 'vacunar', 'verga', 'vergación', 'verijón', 'zuma' and 'ñinga';
  - **"biased words"**[38]: 'afirmó', 'aseguró', 'indicó', 'descubrió', 'reveló', 'supuesto', 'pretendido', 'aunque', 'naturalmente', 'evidentemente', 'indudablemente', 'obviamente', 'claramente', 'indiscutiblemente', 'principalmente', 'básicamente', 'especialmente', 'irónicamente', 'sorprendentemente', 'desafortunadamente', 'afortunadamente', 'curiosamente', 'tristemente', 'trágicamente', 'escándalo', 'polémica', 'controversia', 'legendario', 'mito', 'teóricamente', 'fundamentalista' and 'secta'.

Bibliography

[1] History pages, February 2007. Available from: http://c2.com/cgi/wiki?HistoryPages [cited 2012-06-21].

[2] Edit copy, April 2012. Available from: http://c2.com/cgi/wiki?EditCopy [cited 2012-06-21].

[3] Douglas Noel Adams. H2G2: The hitchhicker's guide to the galaxy, 1979.

[4] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *Proceedings of the 16th international conference on World Wide Web*, WWW '07, pages 261–270, New York, NY, USA, 2007. ACM. doi:10.1145/1242572.1242608.

[5] Maik Anderka and Benno Stein. A breakdown of quality flaws in wikipedia. In *Proceedings of the 2nd Joint WICOW/AIRWeb Workshop on Web Quality*, WebQuality '12, pages 11–18, New York, NY, USA, 2012. ACM. doi:10.1145/2184305.2184309.

[6] Patti Bao, Brent Hecht, Samuel Carton, Mahmood Quaderi, Michael Horn, and Darren Gergle. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, CHI '12, pages 1075–1084, New York, NY, USA, 2012. ACM. doi:10.1145/2207676.2208553.

[7] BBC. Hungary's Pal Schmitt resists quit calls over thesis, March 2012. Available from: http://www.bbc.co.uk/news/world-europe-17561105 [cited 2012-06-23].

[8] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. D3: Data-driven documents. *IEEE Trans. Visualization & Comp. Graphics (Proc. InfoVis)*, 2011. Available from: http://vis.stanford.edu/papers/d3.

[9] U. Brandes and J. Lerner. Visual analysis of controversy in user-generated encyclopedias. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 179 –186, 30 2007-nov. 1 2007. Available from: http://ieeexplore.ieee.org/search/srchabstract.jsp?arnumber=4389012, doi:10.1109/VAST.2007.4389012.

[10] J. Brooke. SUS — A quick and dirty usability scale. *Usability evaluation in industry*, 189:194, 1996. Available from: http://www.usabilitynet.org/trump/documents/Suschapt.doc.

[11] F. Chevalier, S. Huot, and J.-D. Fekete. Wikipediaviz: Conveying article quality for casual wikipedia readers. In *Pacific Visualization Symposium (PacificVis), 2010 IEEE*, pages 49 –56, march 2010. doi:10.1109/PACIFICVIS.2010.5429611.

[12] Pedro Passos Coelho. Candidato do PSD à presidência da assembleia da república, April 2011. Available from: https://www.facebook.com/notes/pedro-passos-coelho/candidato-do-psd-%C3%A0-presid%C3%AAncia-da-assembleia-da-rep%C3%BAblica/10150153318796246.

[13] Stephen Colbert. The Colbert Report — The Word: Wikiality, July 2006. Available from: http://www.colbertnation.com/the-colbert-report-videos/72347/july-31-2006/the-word---wikiality.

[14] Ward Cunningham. Wiki wiki web, 1995. Available from: http://c2.com/cgi/wiki?WikiWikiWeb.

[15] Cory Doctorow. Wikipedia's list of common misconceptions, January 2011. Available from: http://www.boingboing.net/2011/01/11/wikipedias-list-of-c.html.

[16] Python Software Foundation. Python v2.7.3 documentation, June 2012. Available from: http://docs.python.org/index.html.

[17] Chris Harrison. clusterball, July 2008. Available from: http://www.chrisharrison.net/projects/clusterball/index.html [cited 2011-08-05].

[18] Chris Harrison. Wikiviz, September 2009. Available from: http://www.chrisharrison.net/projects/wikiviz/index.html [cited 2011-08-05].

[19] S. Havre, E. Hetzler, P. Whitney, and L. Nowell. Themeriver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions on*, 8(1):9–20, jan/mar 2002. Available from: http://ieeexplore.ieee.org/search/srchabstract.jsp?arnumber=981848, doi:10.1109/2945.981848.

[20] Todd Holloway, Miran Bozicevic, and Katy Börner. Analyzing and visualizing the semantic coverage of wikipedia and its authors: Research articles. *Complex.*, 12:30–40, January 2007. Available from: http://dl.acm.org/citation.cfm?id=1210555.1210561, doi:10.1002/cplx.v12:3.

[21] MediaWiki. Api:etiquette — mediawiki, the free wiki engine, 2012. Available from: http://www.mediawiki.org/w/index.php?title=API:Etiquette&oldid=535181 [cited 2012-06-1].

[22] Santiago M. Mola-Velasco. Wikipedia Vandalism Detection Through Machine Learning: Feature Review and New Proposals - Lab Report for PAN at CLEF 2010. In Martin Braschler, Donna Harman, and Emanuele Pianta, editors, *Notebook Papers of CLEF 2010 LABs and Workshops*, Padua, Italy, September 2010. Available from: http://www.clef2010.org/resources/proceedings/clef2010labs_submission_25.pdf.

[23] Randall Munroe. XKCD #843: Misconceptions, January 2011. Available from: http://www.xkcd.com/843/.

[24] Ryan North. Everytopicintheuniverseexceptchickens dot com: Save wikipedia! promote accuracy at the expense of chickens.:, November 2006. Available from: http://www.everytopicintheuniverseexceptchickens.com/.

[25] Martin Potthast, Benno Stein, and Robert Gerling. Automatic vandalism detection in wikipedia. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen White, editors, *Advances in Information Retrieval*, volume 4956 of *Lecture Notes in Computer Science*, pages 663–668. Springer Berlin / Heidelberg, 2008. Available from: http://www.uni-weimar.de/medien/webis/publications/papers/stein_2008c.pdf, doi:10.1007/978-3-540-78646-7_75.

[26] Nathalie Henry Riche, Bongshin Lee, and Fanny Chevalier. ichase: supporting exploration and awareness of editing activities on wikipedia. In *Proceedings of the International Conference on Advanced Visual Interfaces*, AVI '10, pages 59–66, New York, NY, USA, 2010. ACM. doi:10.1145/1842993.1843004.

[27] Sofia Rodrigues and Nuno Simas. Fernando Nobre renuncia ao mandato de deputado, July 2011. Available from: http://publico.pt/1501403.

[28] Peter Schonhofen. Identifying document topics using the wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, WI '06, pages 456–462, Washington, DC, USA, 2006. IEEE Computer Society. doi:10.1109/WI.2006.92.

[29] Bongwon Suh, Ed H. Chi, Aniket Kittur, and Bryan A . Pendleton. Lifting the veil: improving accountability and social transparency in wikipedia with wikidashboard. In *Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, CHI '08, pages 1037–1040, New York, NY, USA, 2008. ACM. doi:10.1145/1357054.1357214.

[30] Bongwon Suh, E.H. Chi, B.A. Pendleton, and A. Kittur. Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 163 –170, 30 2007-nov. 1 2007. Available from: http://ieeexplore.ieee.org/search/srchabstract.jsp?arnumber=4389010, doi:10.1109/VAST.2007.4389010.

[31] Fernanda B. Viégas and Martin Wattenberg. history flow: visualizing the editing history of wikipedia pages, August 2006. Available from: http://www.research.ibm.com/visual/projects/history_flow/ [cited 2011-08-05].

[32] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations, April 2004. Available from: http://alumni.media.mit.edu/~fviegas/papers/history_flow.pdf [cited 2011-08-05].

[33] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, and Hady Wirawan Lauw. On ranking controversies in wikipedia: models and evaluation. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 171–182, New York, NY, USA, 2008. ACM. doi:10.1145/1341531.1341556.

[34] Howard T. Welser, Dan Cosley, Gueorgi Kossinets, Austin Lin, Fedor Dokshin, Geri Gay, and Marc Smith. Finding social roles in wikipedia. In *Proceedings of the 2011 iConference*, iConference '11, pages 122–129, New York, NY, USA, 2011. ACM. doi:10.1145/1940761.1940778.

[35] Wikcionario. Categoría:es:términos malsonantes — wikcionario, el diccionario libre, 2012. Available from: http://es.wiktionary.org/w/index.php?title=Categor%C3%ADa:ES:T%C3%A9rminos_malsonantes&oldid=721973 [cited 2012-06-13].

[36] Wikcionário. Categoria:obscenidade (português) — wikcionário, o dicionário livre, 2011. Available from: http://pt.wiktionary.org/w/index.php?title=Categoria:Obscenidade_(Portugu%C3%AAs)&oldid=1077460 [cited 2012-06-13].

[37] Wikimedia. API:Main page, June 2011. Available from: http://www.mediawiki.org/w/index.php?title=API:Main_page&oldid=407852 [cited 2011-08-04].

[38] Wikipedia. Wikipedia:palabras que evitar — wikipedia, la enciclopedia libre, 2011. Available from: http://es.wikipedia.org/w/index.php?title=Wikipedia:Palabras_que_evitar&oldid=51594912 [cited 2012-06-13].

[39] Wikipedia. Category:profanity — wikipedia, the free encyclopedia, 2012. Available from: http://en.wikipedia.org/w/index.php?title=Category:Profanity&oldid=492828579 [cited 2012-05-27].

[40] Wikipedia. List of countries where portuguese is an official language — wikipedia, the free encyclopedia, 2012. Available from: http://en.wikipedia.org/w/index.php?title=List_of_countries_where_Portuguese_is_an_official_language&oldid=498321209 [cited 2012-06-23].

[41] Wikipedia. Wikipedia:manual of style (words to watch) — wikipedia, the free encyclopedia, 2012. Available from: http://en.wikipedia.org/w/index.php?title=Wikipedia:Manual_of_Style_(words_to_watch)&oldid=493088522 [cited 2012-05-27].

[42] Wikipedia. Wikipedia:protection policy — wikipedia, the free encyclopedia, 2012. Available from: http://en.wikipedia.org/w/index.php?title=Wikipedia:Protection_policy&oldid=498543950 [cited 2012-06-23].

[43] Wikipédia. Catégorie:insulte — wikipédia, l'encyclopédie libre, 2012. Available from: http://fr.wikipedia.org/w/index.php?title=Cat%C3%A9gorie:Insulte&oldid=78742006 [cited 2012-06-13].

[44] Wikipédia. Wikipédia:palavras a se tomar cuidado — wikipédia, a enciclopédia livre, 2012. Available from: http://pt.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:Palavras_a_se_tomar_cuidado&oldid=30632405 [cited 2012-06-13].

[45] Wikipédia. Wikipédia:termes à utiliser avec précaution — wikipédia, l'encyclopédie libre, 2012. Available from: http://fr.wikipedia.org/w/index.php?title=Wikip%C3%A9dia:Termes_%C3%A0_utiliser_avec_pr%C3%A9caution&oldid=78442332 [cited 2012-06-13].

[46] Jamie Wilkinson. Wikiswarm: visualize wikipedia page histories, January 2009. Available from: http://jamiedubs.com/wikiswarm-visualize-wikipedia-page-histories [cited 2011-08-04].