

Structuring Wiki Revision History*

Mikalai Sabel

University of Trento

msabel@dit.unitn.it

Abstract

Revision history of a wiki page is traditionally maintained as a linear chronological sequence. We propose to represent revision history as a tree of versions. Every edge in the tree is given a weight, called *adoption coefficient*, indicating similarity between the two corresponding page versions. The same coefficients are used to build the tree. In the implementation described, adoption coefficients are derived from comparing texts of the versions, similarly to computing edit distance. The tree structure reflects actual evolution of page content, revealing reverts, vandalism, and edit wars, which is demonstrated on Wikipedia examples. The tree representation is useful for both human editors and automated algorithms, including trust and reputation schemes for wiki.

Categories and Subject Descriptors H.5.3 [Information Interfaces and Presentation]: Collaborative Computing—computer-supported cooperative work, web-based interaction; I.7.1 [Document and Text Processing]: Document and Text Editing—Version control

General Terms Algorithms, design, experimentation

Keywords Wikipedia, revision history, wiki

1. Introduction

Wiki is a thriving web-technology, invented in 1995 [3] and experiencing steady increase of popularity in recent years [4]. Wiki is a site edited directly in browser by visitors. The editing process is simplified and requires very little skills. Wiki is receiving attention from researchers as a prominent and challenging tool for online communities.

In this work, we describe method of organizing versions of a wiki page into a tree, instead of conventional chronolog-

ical sequence. First, a quantitative measure of similarity between individual versions is introduced. This similarity measure, *adoption coefficients*, is defined according to community context, available computation power, *etc.* In this work, we employ a simple yet reasonable automated way to calculate adoption coefficients, by comparing texts of the versions and counting number of inserted and deleted blocks, similarly to edit/Levenshtein distance [7]. Analogous approaches to quantify changes between versions within wiki are used in [1] and [11].

Our major contribution to wiki research is the mechanism to arrange versions of wiki page history into a weighted tree structure that reflects the actual page evolution. We believe that this mechanism will be of great use for emerging wiki reputation, trust and reliability algorithms, e.g. [11, 8]. Secondly, adoption coefficients themselves and their visualization is a powerful tool for wiki analysis, complementary to HistoryFlow [9]. HistoryFlow focuses on the detailed evolution of text fragments, while adoption coefficients and version trees provide general view of content evolution.

Related work about wiki history is performed in the context of trust and reputations, see [11], [5], and [1]. Among those, [11] strictly follows chronological sequence of versions, assuming that trustworthiness of each version depends only on previous version, author reputation and amount of changes introduced. On the other hand, [1] uses the similarity coefficients to update reputations for all previous versions in a fixed range (3 or 10 previous versions, as a rough guide). Tree structure allows variable number of connected versions, and is not limited to chronological neighbors. [5], [1] and [9] focus on tracking individual fragments of text within a page, while this work and [11] address pages as a whole. There are versioning systems that support tree history visualization (e.g. wiki in the recently released CURE system [6], or CVS management systems), but a common approach is to allow users to define parents for the newly entered versions. Our approach automates this task and prevents possible malicious manipulations on the history structure.

We demonstrate and evaluate our approach by applying it to real wiki pages extracted from Wikipedia [10]. Wikipedia is a popular online encyclopedia edited by open online community [4]. It is the most successful example of existing

*This paper is based upon work supported by *Provincia Autonoma di Trento* through the *WILMA Project*.

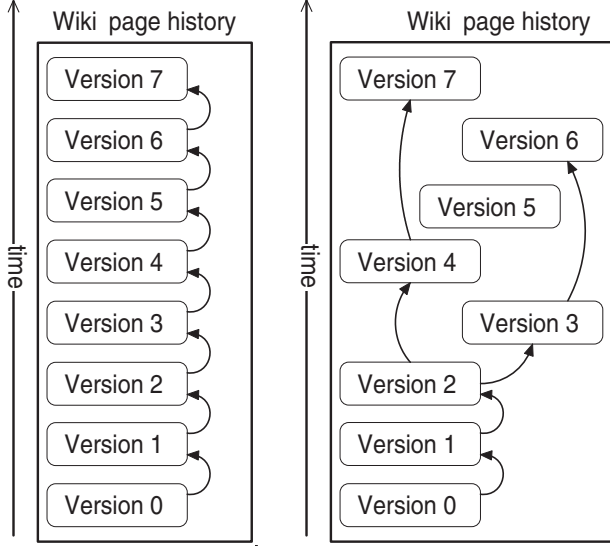


Figure 1. Linear (left) and tree (right) representations of wiki page history

wiki, today Wikipedia is the 10th top site in the Internet (according to *alexa.com* statistics).

2. Wiki page as a tree of versions

Wiki page history is commonly represented as a linear sequence of versions, ordered by edition time (Figure 1, left). However, the actual development of a wiki page is not strictly monotonic. For example, malicious or non-conforming edits are corrected by reverting to previous versions. In case of a disputable page, a sequence of back-and-forth edits may happen, called *edit war*. All these scenarios have been observed in Wikipedia [1, 9]. Even during a peaceful content evolution, parts of added information may lose importance and be removed.

To model evolution of a wiki content better, we propose to use a directed tree graph (Figure 1, right). Each version has at most one *parent* version. The parent is considered to be the major foundation for its child version(s). The very first version and versions re-written from scratch have no parents. (Thus the graph becomes formally not a tree, but a forest. This inconvenience is evaded by introducing additional edges with zero weights.) The tree model is reasonably simple and captures typical behavior of a wiki editor: a new version is normally created starting from one of the existing versions, but not necessarily the latest.

Each edge in the tree has a weight representing the degree of similarity between the corresponding versions. The weight is called *adoption coefficient*, and it estimates how much of the parent version’s content is reused in the child version.

Adoption coefficients actually provide the way to build the tree from initial sequence of page versions. First, for

a given wiki page, adoption coefficients are calculated for all pairs of versions. Next, for each version, the parent is chosen among previous versions as the version having the maximum adoption coefficient. Finally, only the coefficients corresponding to the tree edges remain.

Generally, any automated algorithm for producing adoption coefficients is feasible, starting from naive text comparison and up to semantic-aware tools, depending on requirements, resources and definition of ‘similarity’ in particular environment. We have implemented a naive text comparison as explained in the next Section.

3. Adoption coefficients

Similarity between page versions are characterized numerically by *adoption coefficients*. Consider two versions i and j of the same page, i being older than j . The *adoption coefficient* $a_{i,j}$ characterizes the ‘similarity’ between the versions, in particular, it measures how much content of version i is preserved in version j . The scale for adoption coefficient is $[0, 1]$, with 0 corresponding to independent versions, and 1 meaning that j is a copy of i , with no differences.

We use simple block-oriented text comparison to measure similarities and differences between page versions, similar approach is adopted in [11, 1];

Adoption coefficients are calculated in the following way:

1. Texts of the two versions to compare are divided into *blocks*, using as separators punctuation marks (full stops, colons, semicolons, exclamation and question marks) and some wiki-specific marks (asterisks (*) and number sign (#)). Consecutive spaces and new lines are ignored, and wiki and HTML mark-up is partially cleared.
2. Size of the second text in *blocks* is measured ($N_{total,j}$).
3. Number of deleted and added *blocks*, $N_{deleted}$ and $N_{inserted}$ respectively, are calculated using standard GNU *diff* program.
4. A normalized adoption coefficient is found as:

$$a_{i,j} = 1 - \frac{N_{inserted} + N_{deleted}}{N_{total,j} + N_{deleted}}, \quad (1)$$

Note that these adoption coefficients are *symmetric*, because $a_{i,j} = a_{j,i}$ always holds. Generally, this property is not necessary.

4. Analyzing Wikipedia examples

In this Section we apply our tools to several wiki pages retrieved from Wikipedia [10]. We demonstrate that regular wiki pages exhibit complex structure, and that the proposed tree representation is able to capture it. Table 1 lists the pages we consider, number of versions retrieved, and creation dates for first and last versions. We intentionally have chosen pages considered in [9, 11], in order to compare our results with previous ones, when possible.

Page Title	Retrieved versions
U.S. National Forest	61 (03/03/2002 – 23/04/2007)
Chocolate	500 (19/12/2001 – 20/05/2005)
Microsoft	500 (03/11/2001 – 10/05/2004)

Table 1. Examined Wikipedia pages

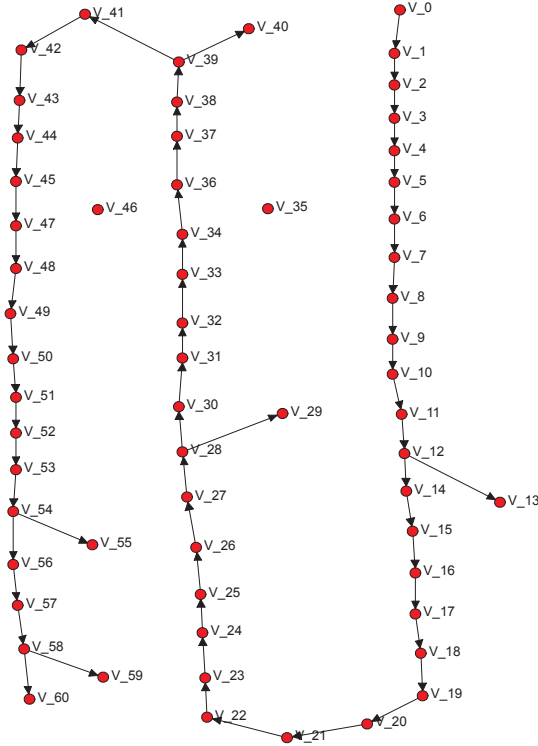


Figure 2. Version tree for "U.S. National Forest"

Below, for each of the pages, we present a tree of versions, calculated by the algorithm described in Sections 2 and 3 and plotted using NetDraw tool [2]. We examine the trees and attempt to explain significant anomalies, such as not-connected versions and branchings. To do that, we consider original comments in the Wikipedia history, and content of the corresponding pages.

4.1 'U.S. National Forest' page

On the version tree (Figure 2) we see that versions 35 and 46 stand apart. In fact they are examples of sheer vandalism, when all page content is replaced with a short unrelated note.

Versions 13, 29, 40, 55 and 59 are not in the main branch. From the comments and original texts in the Wikipedia we conclude that:

- version 13 is one of thee consecutive edits by the same author, with minor changes of the same fragment. It is not clear if the author intentionally revoked some of his changes, or if it was just a coincidence. To avoid such

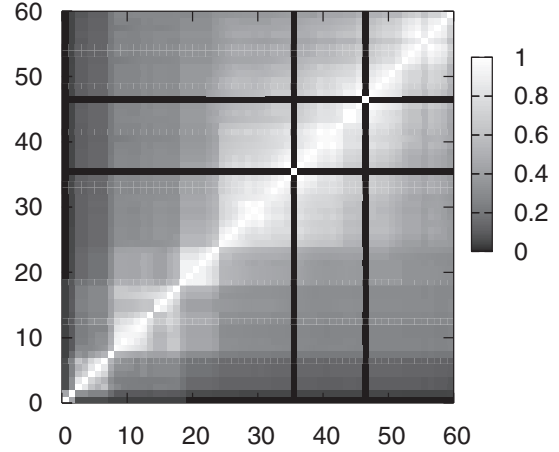


Figure 3. Map of adoption coefficients for 'U.S. National Forest' page

confusions, we suggest to skip consecutive versions from the same author, leaving only the last of them, as in [1];

- version 29 was reverted, because an anonymous editor added an image, which presumably did not have appropriate copyright (the image itself is not available now);
- version 40 was reverted, because an anonymous editor added a link to site that was considered inappropriate by another editor;
- version 55 was a vandalism by an anonymous editor: insertion of a small piece of offensive text into the page;
- in version 59 was a vandalism by an anonymous editor: a paragraph of text was deleted, and then restored in version 60;

Remarkably, no other malicious edits were mentioned in the history comments. All edits marked 'remove' or 'rv' are mentioned above. This implies that in this case our algorithm detected all significant inappropriate or malicious edits.

In addition to the version tree, Figure 3 presents an images displaying whole matrix of *adoption coefficients* $A = \{a_{i,j}\}$. Brightness of each point corresponds to the value of $a_{i,j}$. Only lower-right half of the picture is actually informative. The upper-left triangle represents a 'virtual inheritance' from the future versions to the past ones. Because of the symmetry of *adoption coefficients*, the image is in fact diagonally-symmetric.

On this picture sheer vandalism appears as distinguishable black crosses around white center points, because vandalized versions are completely different from all others. White 'squares' of different intensity and size correspond to stable periods in page development, when the content does not change significantly between all involved versions. Brighter blocks mean less changes happened within.

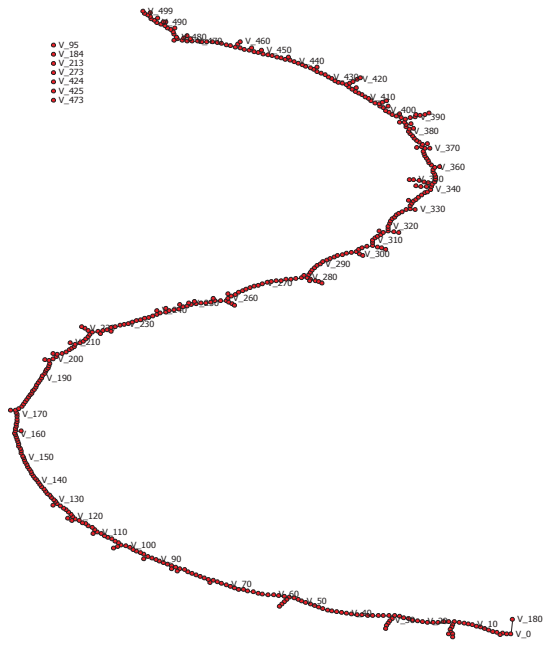


Figure 4. First 500 versions of page "Chocolate"

4.2 'Chocolate' page, early versions

Figure 4 presents version tree for the first 500 versions of Wikipedia page 'Chocolate'. Although the main 'backbone' is apparent, the detailed structure turned out too complex for a complete explanation. Instead of this large tree, we address only a part of it, shown on Figure 5.

Notable is branching around versions 26 — 37. It corresponds to an *edit war* between two authors reported previously in [9].

The strange branch of versions 135, 215, 452, ..., coming out from node 14 is, in fact, a set of vandalism cases (deletions of significant parts of the article).

The branch around versions 53-58 is important. A new paragraph was introduced through versions 54-57, presumably by the same author. Then another editor added few changes to the page, including rephrasing the new paragraph. In this case, our text-based metric was unable to detect that version 58 contains significant information from version 57.

Version 95 is a sheer vandalism (mass deletion). Versions 83 and 86 and parts of consecutive edits from the same users, and version 74 deals with re-formatting tables and images in the page.

4.3 'Microsoft' page, early versions

Again, the tree of first 500 versions (Figure 6) is too large for a detailed examinations, and we focus only on a part of it (Figure 7).

Version 92 is a mass deletion vandalism. Versions 13 and 75 — 76 are intermediate edits in sequences of edits by the same authors.

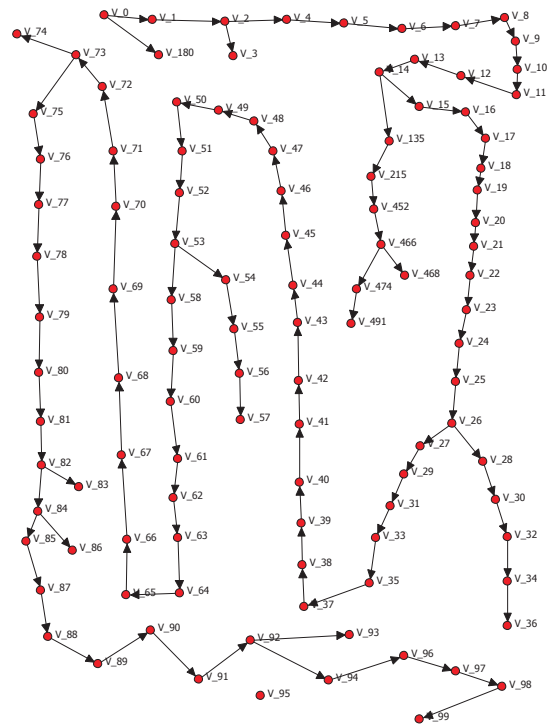


Figure 5. Tree fragment for page "Chocolate"

Version 21 was reverted, because it was a biased and non-appropriate modification.

The branch of versions 89 — 94 actually provides a useful content. It is disconnected from the main branch, because version 95 provided a large set of minor clarity edits throughout the text. According to adoption coefficients, it differs significantly from all previous versions, and version 88 turned out to be the closest match. This is a fault of our tool. It demonstrates that numerous simultaneous 'clarity edits' are hard to handle given the coarse-grained block-based text comparison.

5. Future work

Currently, we are looking for a suitable tool to visualize version trees, allowing to shape the graph according to version timestamps, and in other specific modes. We also consider several improvements to our Wikipedia-related tools. First, skipping subsequent versions from the same author, in order to avoid dealing with incomplete edits. Second, a better support for wiki- and HTML- syntax should improve accuracy of adoption coefficients calculations. Current Wikipedia articles often include complex formatting, tables, images, and frames that act as additional 'text noise' to our algorithm. After that, a wider and deeper tests on Wikipedia pages will become possible.

Adoption coefficients and version trees are employed in our reputation scheme for wiki [8]. The scheme is still under development, and reducing errors in generated trees is an im-

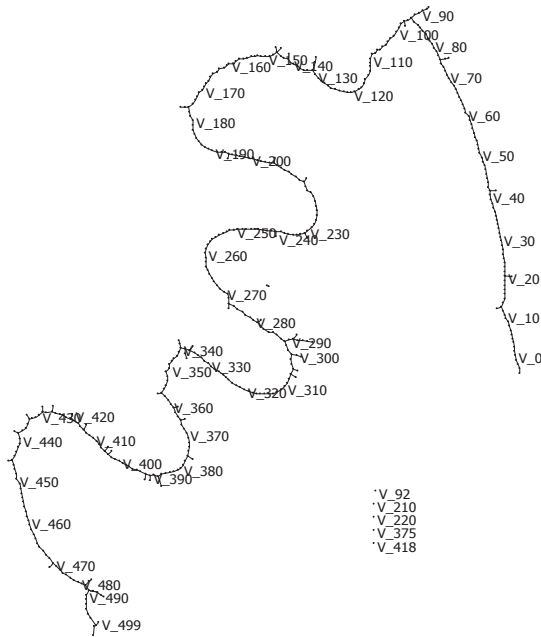


Figure 6. First 500 versions of page "Microsoft"

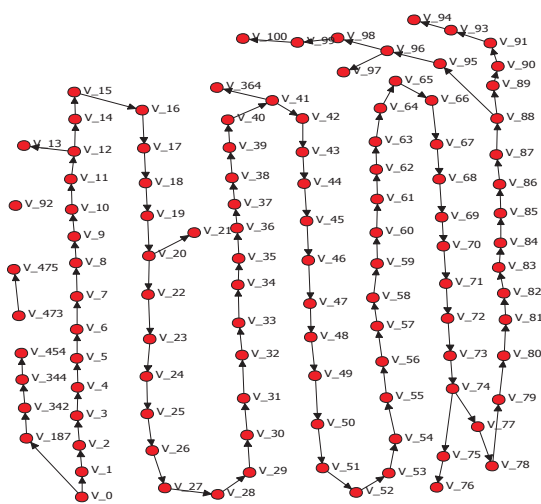


Figure 7. Tree fragment for page "Microsoft"

portant goal. Some of already developed reputation and trust schemes for wiki can benefit from using tree version structure instead of the linear one. For example, if the approach of [11] is applied to main branch only, it will avoid 'gaps' in text evolution created by mass deletions.

6. Conclusions

In this work, we present a new approach to interpreting revision history in wiki. Saved versions are compared to each other, the most similar versions assume parent-child relations and form a tree of versions. In the process, each edge of the tree receives a weight, called *adoption coefficient*,

that measures how much parent's content is presumed in the child. This allows to automatically create structured page history that follows the actual page content evolution. Users are not required and, in fact, *cannot* manipulate structure of the version tree without modifying page content, which is eminently suitable for wiki-like systems.

We described implementation of a tool for building revision history trees. Adoption coefficients are calculated similarly to edit distance between versions' texts.

We demonstrate performance of our tool on 'real-life' Wikipedia pages, and indicate discovered advantages and shortcomings. In most cases, the proposed approach succeeds in following non-linear wiki content evolution. All code and data used in this paper are available online at <http://dit.unitn.it/~msabel/wikisym2007/>

References

- [1] T. B. Adler and L. de Alfaro. A content-driven reputation system for the Wikipedia. In *WWW2007*, May 2007.
- [2] S. Borgatti. NetDraw: Graph visualization software. Harvard: Analytic Technologies, 2002.
- [3] W. Cuningham and contributing authors. <http://c2.com/cgi/wiki>.
- [4] A. Lih. Wikipedia as participatory journalism: Reliable sources? Metrics for evaluating collaborative media as a news resource. In *Proceedings of the International Symposium on Online Journalism*, 2004.
- [5] D. L. McGuinness, H. Zeng, P. P. da Silva, L. Ding, D. Narayanan, and M. Bhaawal. Investigations into Trust for Collaborative Information Repositories: A Wikipedia Case Study. In *Proceedings of the Workshop on Models of Trust for the Web*, May 2006.
- [6] T. Schümmer. CURE project. <http://cure.sourceforge.net/>.
- [7] G. Navarro. A guided tour to approximate string matching. *ACM Comput. Surv.*, 33(1):31–88, 2001.
- [8] M. Sabel, A. Garg, and R. Battiti. WikiRep: Digital reputations in virtual communities. In *Proceedings of XLIII Congresso Annuale AICA*, pages 209–217, Oct 2005.
- [9] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 575–582, New York, NY, USA, 2004. ACM Press.
- [10] Wikipedia, the free encyclopedia. wikipedia.org.
- [11] H. Zeng, M. Alhossaini, L. Ding, R. Fikes, and D. L. McGuinness. Computing Trust from Revision History. In *Proceedings of the 2006 International Conference on Privacy, Security and Trust*, October 2006.