The Analysis and Visualization of Entries in Wiki Services

Jakub Gawryjołek and Piotr Gawrysiak

Institute of Computer Science, Warsaw University of Technology
Nowowiejska 15/19, 00-665 Warsaw, Poland
{J.Gawryjolek, P.Gawrysiak}@elka.pw.edu.pl

Summary. The use of online collaboration environments has become exceptionally widespread over the past decade. One of the most popular styles of collaboration are the "wiki" web sites. They have attracted attention because of their policy of letting anyone become an editor. This paper presents the technique for the analysis and visualization of Wikipedia - the largest wiki in existence. Specifically, it concentrates on some activity patterns of its contributors. First, a new visualization and analysis tool named JWikiVis is presented. Second, with the use of this software, some interesting user behaviors are described. Finally, text classification algorithms are applied in order to determine some patterns observed in individual wiki pages as well as in the entire service.

Key words: information visualization, Wiki, collaboration, text categorization, text and web mining

1 Introduction

A wiki is a type of Web site that gives every user the possibility to contribute to its content, very often without the need for registration. Such an approach makes a wiki an effective tool for collaborative authoring. Vandalism seems as a natural consequence of the open philosophy of this technology. The only outcome that we should expect is mess and vulgarism on the pages. And these things happen very often, still wikis seem to work very well. The important question is who does what and what constitutes the success of wiki technology? Furthermore, exploring multiple visual presentations, or visualizations, often helps a user make sense of a large collection [9].

Although most wikis are open to the public, an organization of roles (readers, editors, reviewers, destroyers, etc.)[1], as defined in Nupedia project, is also visible. Most of the actions like creations, mass and small deletions, corrections, and swapping of some parts of text, when visualized and analyzed, can provide useful information concerning behaviors of groups, for instance the discrepancies between anonymous and registered users. Furthermore, the statistical analysis of articles together with the graphical representation can serve as a tool for finding some interesting trends within the service or for the prediction of the direction of a wiki evolution. Obviously the most interesting wiki for research is Wikipedia, with its 1 500 000 articles in English version only and multilingual content (see [11]).

The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 explains visualization approach and algorithms used. Section 4 presents analysis of behaviors and trends observed in the system. Finally, section 5 describes possible future development directions and summarizes the paper.

2 Review of the Related Work

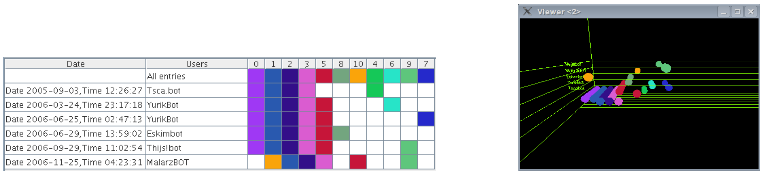
The philosophy of most popular wiki system - MediaWiki - is that it should facilitate correction of mistakes, rather than preventing them. MediaWiki's "diff" and "hist" features are such tools that help in restoring the article's content, viewing changes etc. They does not allow, however, to depict whole development history of the article. The research conducted by F. Viegas et al. [4] shows how such broader system - called History Flow - might be constructed, using highly visual means. Our approach, dubbed JWikiVis, has been highly inspired by History Flow, and delivers similar visualization power addressing History Flow deficiencies. Other work in the area is The ThemeRiver visualization[9] which depicts thematic variations over time within a large collection of documents. Visualizing the affective structure of a text document is the subject matter of [5] and graphical representation of interaction in an on-line collaboration environment is described in [2]. Finally, the semantic coverage of Wikipedia and its authors is dealt with in [10].

3 The Visualization System

The main corpus that we used for visualization and analysis was Polish version of Wikipedia, specifically the complete page edit history from November 2006, totaling 520882 pages and 5158509 revisions, resulting in 28GB of data. The visualization software - JWikiVis - has been written mainly in Java, with 3-D components partially implemented in C++.

3.1 JWikiVis - The Visualization Part

True text visualizations should represent textual content and meaning to analysts without them having to read it in the manner that text normally requires[6]. Main factors that determined the type of the visualization was the structure of text in Wikipedia articles. A text is a string of characters which is modified at some intervals. Such an approach limits the illustration possibilities to two dimensions with text-corresponding structures on one axis, and the flow of time on the second. In this fashion works the main part of JWikiVis visualization. The example in figure 1(a) shows the result for the article about Italian county, Jolanda Di Savoia.



Date	Users	0	1	2	3	5	8	10	4	6	9	7
	All entries	🟪	🟦	🟦	🟪	🟥	🟩	🟧	🟩	🟩	🟩	🟦
Date 2005-09-03,Time 12:26:27	Tsca.bot	🟪	🟦	🟦	🟪				🟩			
Date 2006-03-24,Time 23:17:18	YurikBot	🟪	🟦	🟦	🟪	🟥				🟩		
Date 2006-06-25,Time 02:47:13	YurikBot	🟪	🟦	🟦	🟪	🟥						🟦
Date 2006-06-29,Time 13:59:02	Eskimbot	🟪	🟦	🟦	🟪	🟥	🟩					
Date 2006-09-29,Time 11:02:54	Thijs!bot	🟪	🟦	🟦	🟪	🟥					🟩	
Date 2006-11-25,Time 04:23:31	MalarzBOT		🟧	🟦	🟦	🟪		🟥			🟩	

(a) 2-D visualization of the article about 'Jolanda Di Savoia'

(b) 3-D Visualization of the article about 'Jolanda Di Savoia'

Fig. 1. JWikiVis 2-D and 3-D visualization

Every single rectangle in the picture presents the part of text - paragraph or single line. In the example, rectangles correspond to the paragraphs. In figure 1(a) there are 11 paragraphs altogether - first row of the table - numbers 0-10. On the left-hand side we can see the name of the contributors and the date of their revisions. Rows represent the article at a certain point of time indicated by the revision time, and columns - parts of text. Columns are the added parts of texts in all revisions - set of all positive entries. Each part is assigned certain column in this set. When a part is placed in certain revision (row), it takes only the position (column) of a part existing in that revision keeping at the same time the correct order of the whole article's text. In last revision in fig. 1(a) the orange rectangle - first part in text - is placed in the second column - first existing part in that version. JWikiVis is not limited only to 2-D visualization. The three-dimensional representation of the above example is shown in figure 1(b). The third, z-axis, is the user axis, on which every user is placed on the separate level, what is helpful for examining certain user's contributions.

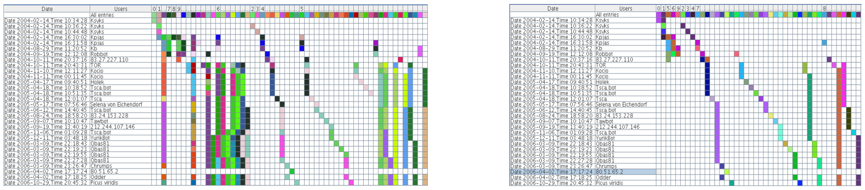
4 The Entries and Trend Analysis

Having all the necessary tools (including document multi-classifier presented in [3]) we can make some analysis of typical contributors' behaviors and trends in Wikipedia. The examination of the behaviors revealed some common patterns like anonymous versus named authorship[3], negotiation, content stability, vandalism and repair described in [4]. Additionally, evolution in subject matter of the documents and the activity of edits in some areas were observed. Here we can present only a small example of the analysis prepared in [3].

4.1 Parts Durability

There are two factors that determine the appearance of JWikiVis's charts. First, the number of parts of individual users. Second, and more important, is the time of existence of these parts. If a user added fewer parts which existed for a long time, their contribution to the chart may be more significant than those who added more, but short-living ones. The full description and visualization from the user perspective is presented in [3].

In order to better understand the durability issue we have to focus on the article as a whole, not only on the individual contributors. It would seem natural that pages would tend to stabilize over time, but pages change in size and turnover in text [4]. Figure 2(a) presents the article about 'Optical Disk'. Notice that the parsing delimiter was a single new line character. Figure 2(b), on the other hand, presents the same article but with a double new line character - a paragraph - as a delimiter.



Date	Users
	All entries
Date 2004-02-14,Time 10:34:28	Ksyks
Date 2004-02-14,Time 10:36:22	Ksyks
Date 2004-02-14,Time 10:44:48	Ksyks
Date 2004-02-14,Time 16:30:02	Kpjas
Date 2004-02-14,Time 16:31:58	Kpjas
Date 2004-08-29,Time 11:20:52	Kb
Date 2004-09-19,Time 12:12:08	Robbot
Date 2004-10-11,Time 20:37:16	83.27.227.110
Date 2004-10-11,Time 20:43:31	TOR
Date 2004-11-10,Time 12:11:27	Kocio
Date 2004-11-11,Time 00:11:45	Kocio
Date 2005-04-17,Time 09:40:51	Holek
Date 2005-04-18,Time 10:38:52	Tsca.bot
Date 2005-04-18,Time 10:51:35	Tsca.bot
Date 2005-04-18,Time 12:01:07	Tsca
Date 2005-05-17,Time 07:56:46	Selena von Eichendorf
Date 2005-06-12,Time 14:40:45	Tsca.bot
Date 2005-08-24,Time 18:58:20	83.24.153.228
Date 2005-09-07,Time 10:10:47	Tawbot
Date 2005-09-19,Time 13:40:19	212.244.107.146
Date 2005-11-06,Time 01:09:28	Tsca.bot
Date 2005-12-11,Time 03:48:18	YurikBot
Date 2006-03-09,Time 22:18:43	Obas81
Date 2006-03-09,Time 22:19:23	Obas81
Date 2006-03-09,Time 22:19:55	Obas81
Date 2006-03-09,Time 22:27:28	Obas81
Date 2006-03-09,Time 23:26:47	Chrumps
Date 2006-04-02,Time 17:17:24	80.51.65.2
Date 2006-04-02,Time 17:18:25	Odder
Date 2006-10-29,Time 20:45:32	Picus viridis

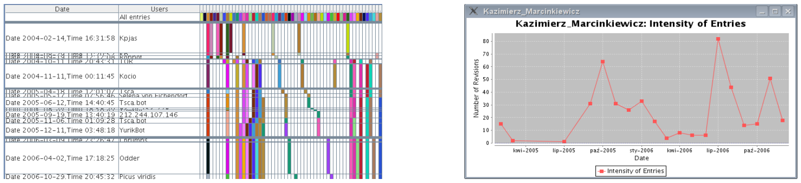
(a) 'Optical Disc' with a single new line character as a delimiter

Date	Users
	All entries
Date 2004-02-14,Time 10:34:28	Ksyks
Date 2004-02-14,Time 10:36:22	Ksyks
Date 2004-02-14,Time 10:44:48	Ksyks
Date 2004-02-14,Time 16:30:02	Kpjas
Date 2004-02-14,Time 16:31:58	Kpjas
Date 2004-08-29,Time 11:20:52	Kb
Date 2004-09-19,Time 12:12:08	Robbot
Date 2004-10-11,Time 20:37:16	83.27.227.110
Date 2004-10-11,Time 20:43:31	TOR
Date 2004-11-10,Time 12:11:27	Kocio
Date 2004-11-11,Time 00:11:45	Kocio
Date 2005-04-17,Time 09:40:51	Holek
Date 2005-04-18,Time 10:38:52	Tsca.bot
Date 2005-04-18,Time 10:51:35	Tsca.bot
Date 2005-04-18,Time 12:01:07	Tsca
Date 2005-05-17,Time 07:56:46	Selena von Eichendorf
Date 2005-06-12,Time 14:40:45	Tsca.bot
Date 2005-08-24,Time 18:58:20	83.24.153.228
Date 2005-09-07,Time 10:10:47	Tawbot
Date 2005-09-19,Time 13:40:19	212.244.107.146
Date 2005-11-06,Time 01:09:28	Tsca.bot
Date 2005-12-11,Time 03:48:18	YurikBot
Date 2006-03-09,Time 22:18:43	Obas81
Date 2006-03-09,Time 22:19:23	Obas81
Date 2006-03-09,Time 22:19:55	Obas81
Date 2006-03-09,Time 22:27:28	Obas81
Date 2006-03-09,Time 23:26:47	Chrumps
Date 2006-04-02,Time 17:17:24	80.51.65.2
Date 2006-04-02,Time 17:18:25	Odder
Date 2006-10-29,Time 20:45:32	Picus viridis

(b) 'Optical Disc' with a double new line character as a delimiter

Fig. 2. Visualization of the same article depending on the delimiter

In fig. 2(a) noticeable is the fact that there are many long columns what indicates that some parts existed through many subsequent revisions. Moreover, initial text of a page usually exists longer and suffers fewer changes than the parts added later. Another indication of that hypothesis can be the "stairs" in the middle part of the picture. Such "stairs" inform that corresponding texts did not last for more than one or two revisions. As the consequence new parts appeared and disappeared very often. The second figure, tells another story. With the exception to three or four parts, none of the paragraphs survived more than 5 revisions. It is because a change in a sentence causes a paragraph to be marked as deleted. Hence, we can conclude that most of the changes were rather small edits within individual lines rather than large modifications of the entire article. Moreover, the graphical representation of Wikipedia articles is very often an unstructured formation. It is because people tend to delete and insert new parts rather than move already existing text. One explanation given by F.Viegeas et al. [4] is that Wikipedia editing window is small what makes it difficult to see whole article at once. Other explanation can be that users usually agree on the order of paragraphs and sentences in the documents, but they think they should be formulated in a different way. There is another issue concerning the durability of parts in the article, namely how long they have existed in time rather than in how many revisions. In other words, we want to see the revisions scaled by date[4], figure 3(a).



Date	Users
	All entries
Date 2004-02-14,Time 16:31:58	Kpjas
[illegible]	[illegible]
[illegible]	Roppor
Date 2004-10-11 Time 20:43:31	TOR
Date 2004-11-11,Time 00:11:45	Kocio
Date 2005-04-18 Time 12:01:07	Tsca
Date 2005-05-17 Time 07:56:46	Selena von Eichendorf
Date 2005-06-12,Time 14:40:45	Tsca.bot
[illegible]	[illegible]
Date 2005-09-19,Time 13:40:19	212.244.107.146
Date 2005-11-06,Time 01:09:28	Tsca.bot
Date 2005-12-11,Time 03:48:18	YurikBot
Date 2006-03-09 Time 23:26:47	Chrumps
Date 2006-04-02,Time 17:18:25	Odder
Date 2006-10-29,Time 20:45:32	Picus viridis

(a) 'Optical Disc' scaled by date

(b) Editing activity of Kazimierz Marcinkiewicz page

Fig. 3. Different visualization of frequency patterns

Some of the parts which existed in many revisions also survived significant amount of time, but there are also some that are almost invisible because of their short time of existence. Specifically, places where users seemed to disagree on the topic are now hidden. This suggests that whenever there is something controversial or users have different opinions on the matter, revisions occur more often. Figure 3(a) is also very helpful in scrutinizing frequency patterns which is the subject matter of the following subsection.

4.2 Frequency Patterns

Users become significantly active in editing certain page in the time of year which somehow corresponds to the content. This behavior is 'marked' by gray horizontal strips in the scaled version of 'Optical Disk' and peaks in figure 3(b) about Polish politician, Kazimierz Marcinkiewicz.By scrutinizing the charts we can find some important facts concerning people's lives, about whom the articles are. Secondly, the anniversaries of important history events is indicated by the increased activity in the corresponding history articles. Finally, high frequency of contributions during the whole period may indicate that the topic is highly controversial. The issue about controversial articles as well as vandalism, disagreement and negotiation patterns are examined in [3].

5 Future Work and Conclusions

One of the most important aspects for the daily use of Wikipedia is the strong interconnection of its articles through the links[7]. Two pages can be treated as neighbors if their links direct to each other or if they link to the third, different page. If two pages point to many same sites, their similarity may increase. In this fashion we could find another measure of similarity of pages and create theme maps [8]. It may be revealing, also, to detect pages having similar visualization structure. We hope that the discrete character of JWikiVis visualization, with some improvements and corrections, may turn out to be a sufficient tool for such analysis.

The evolution of a topic in Wikipedia is a complex and long process during which many patterns occur. Contributors act in different, positive and negative roles. Negotiations, disagreements, acts of vandalism are the examples of possible behaviors taking place in wiki communities. JWikiVis is a visualization software that helps to understand how collaborative documents are created and how they evolve over time. MediaWiki engine together with Wikipedia's Talk pages and forums provide tools to heal undesirable effects. However, in order to understand and possibly prevent these activities as well as to have a detailed insight into many positive patterns a visualization information is needed. We hope that JWikiVis and its future development can satisfy some of these needs.

References

1. L. Aronsson. Operation of a large scale, general purpose wiki website: Experience from susning.nu's first nine months in service. In *Verlag für Wissenschaft und Forschung*, 2002.
2. R. P. Biuk-Aghai. Visualization of interactions in an online collaboration environment. In *Collaborative Technologies and Systems, 2005. Proc. of the 2005 International Symposium*, 2005.
3. J. Gawryjołek. The analysis and visualization of entries in wiki services, 2007. Institute of Computer Science, Warsaw University of Technology, BSc. Thesis.
4. K. Dave F. Viegas, M. Wattenberg. Studying cooperation and conflict between authors with history flow visualizations. In *Proc. of SIGCHI*, 2004.
5. H. Lieberman H. Liu, T. Selker. Visualizing the affective structure of a text document. In *Conference on Human Factors in Computing Systems*, 2003.
6. K. Pennock D. Lantrip M. Pottier A. Schur V. Crow J.A. Wise, J.J. Thomas. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Proc. on Information Visualization*, 1995.
7. M. Völkel M. Krötzsch, D.Vrandecic. Wikipedia and the semantic web - the missing links. In *Proc. of Wikimania 2005 - The First International Wikimedia Conference. Wikimedia Foundation.*
8. M. Brewster H. Foote N. E. Miller, P. C. Wong. Topic islands - a wavelet-based text visualization system. In *IEEE Visualization, Proc. of the Conference on Visualization*, 98.
9. P. Whitney L. Nowell S. Havre, E. Hetzler. Themeriver: visualizing thematic changes in large document collections. *Visualization and Computer Graphics, IEEE Transactions*, Jan/Mar 2002.
10. K. Börner T. Holloway, M. Božičević. Analyzing and visualizing the semantic coverage of wikipedia and its authors. In *Comlexity, Special issue on Understanding Complex Systems.*
11. C. S. Ang U. Pfeil, P. Zaphiris. Cultural differences in collaborative authoring of wikipedia. *Journal of Computer-Mediated Communication,12(1),art. 5*, 2006.