WikiChanges - Exposing Wikipedia Revision Activity

Sérgio Nunes[1]
sergio.nunes@fe.up.pt

Cristina Ribeiro[1,2]
mcr@fe.up.pt

Gabriel David[1,2]
gtd@fe.up.pt

[1]Departamento de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto
[2]INESC-Porto
Rua Dr. Roberto Frias, s/n
4200-465 Porto, Portugal

ABSTRACT

Wikis are popular tools commonly used to support distributed collaborative work. Wikis can be seen as virtual scrapbooks that anyone can edit without having any specific technical know-how. The Wikipedia is a flagship example of a real-word application of wikis. Due to the large scale of Wikipedia it's difficult to easily grasp much of the information that is stored in this wiki. We address one particular aspect of this issue by looking at the revision history of each article. Plotting the revision activity in a timeline we expose the complete article's history in a easily understandable format. We present WIKICHANGES, a web-based application designed to plot an article's revision timeline in real time. WIKICHANGES also includes a web browser extension that incorporates activity sparklines in the real Wikipedia. Finally, we introduce a revisions summarization task that addresses the need to understand what occurred during a given set of revisions. We present a first approach to this task using tag clouds to present the revisions made.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.3 [**Group and Organization Interfaces**]: Web-based interaction; H.4.3 [**Information Systems Applications**]: Communications Applications—*information browsers*

General Terms

Experimentation, Human Factors

1. INTRODUCTION

Wikis are unique software systems that allow users to easily create and maintain web documents. The most distinctive feature of a wiki is the ease with which web pages can be created or updated. Typically, no authentication is needed to perform actions on a wiki system, making them ideal tools for online collaborative information systems. Another important feature of wikis that is present in most implementations is the preservation of the entire revision history of every web page. This feature is useful for undoing revisions but may also be useful for viewing the content's evolution.

Wikis have become a disruptive technology that has led to the development of new products mostly in the field of information management. The Wikipedia[1] is a prime example of a service made possible by the use of wikis. Wikipedia is an open web-based encyclopedia developed on top of the free MediaWiki software. Each article is the result of the collaboration of many volunteers from around the world. Currently (April 2008), Wikipedia has more than 10 million articles written in 252 languages and is ranked among the top ten most-visited sites in the word [1]. The English Wikipedia alone has more than 2 million articles.

Since the complete revision history of each single article is preserved, there is an enormous quantity of information available from Wikipedia's past. We explore this information by looking at the number of revisions over time for each article. In other words, we observe the distribution of revisions through time by plotting historical profiles of Wikipedia articles. We show that in high-visibility topics the distribution of the revision activity over time is highly correlated with the popularity of a given subject over the same period. Following these findings we identify an application for the task of *changes summarization* and propose a simple approach based on tag clouds. These features were incorporated in a fully functional prototype named WIKICHANGES.

This paper is organized as detailed next. In Section 2 we describe the construction of a timeline for a Wikipedia article and present examples that illustrate the valuable information captured in these timelines. In Section 3 we introduce the problem of changes summarization and present a simple approach. In Section 4 we present the prototype that we have developed to explore, compare and analyze revision timelines. We also present a client-side extension to Wikipedia that embeds the timeline next to the article's title. A survey of related work is included in Section 5. In Section 6 we present our main conclusions and future work ideas.

2. WIKIPEDIA REVISION ACTIVITY

The *update profile* of an article is the distribution over time of the revisions made to that article in a wiki system. We consider all changes made to an article, also including

WikiSym '08, September 8-10, Porto, Portugal.
Copyright 2008 ACM 978-1-60558-128-3/08/09 ...$5.00.

[1] http://www.wikipedia.org

changes that are classified as SPAM. Figure 1 show two side by side plots of the complete revision history of Wikipedia's articles *2005* and *2007*. These articles are hubs that point to the events reported in Wikipedia each year. The number of events being reported yearly is growing fast in recent years. The article *2005* had 5,714 revisions, while the article *2007* had 11,494 revisions.

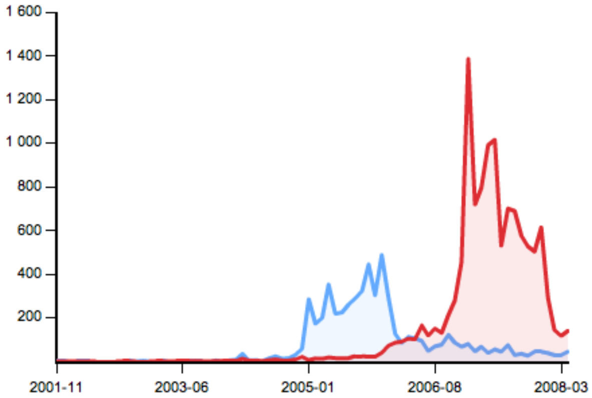


Figure 1: Revision history plots for Wikipedia articles *2005* (brighter color) and *2007* (darker).

Building an update profile for a Wikipedia article is a simple process. For a given concept expressed as a string, we first perform a disambiguation step where we match the string to a specific Wikipedia article. Then, using the Wikipedia API[2] we iterate through all revisions of the article and extract each revision's date. Finally, we count the number of revisions by period (e.g. day, month, year) and produce a standard time series. Due to the existence of missing values, we need to fill the time series with empty "slots" to have an uniformly sampled series. To illustrate this, consider an article that has been created in January and then updated in March. To get a consistent time series we need to automatically insert February with zero updates.

We've built update profiles for several pages based on popular subjects. We found a striking correlation between the popularity of the topic (as perceived by mainstream media) and its profile. The number of updates to the article grows significantly while the event is taking place, both due to the flow of new information and to the increased attention given to the article. The update profile for the article *Tour de France* (Figure 4) is an example of a recurring, predictable event clearly captured in the revision activity timeline.

In Figure 3 the reaction to an unexpected events, specifically the disappearance of *Steve Fossett*, is shown. A very significant peak occurred in the article in September 2007, when the adventurer was reported missing during a flight. Later, a smaller burst occurred in February 2008 when he was declared death after months of searches.

We found similar patterns by manually inspecting more than 80 individual Wikipedia articles about unexpected events. Even when the topic is only of regional interest (e.g. Portuguese presidential elections), thus containing fewer revisions, activity bursts are clearly identifiable.

3. REVISIONS SUMMARIZATION

When we know in advance the events associated with a given article, bursts in the revision timeline are predictable and understandable. However, often we are not aware of all details of a given topic and some bursts might come as a surprise. When used in an exploratory way these timelines are poor at providing insights to specific bursts in activity. Considering this, we decided to explore the automatic construction of summaries for a given period.

Given a set of sequential revisions to a specific document (typically spanning a given period), our goal is to produce an automatic summary of the revisions made. In the context of Wikipedia the applications are obvious, specifically for understanding bursts in the revision activity of articles or tracking changes for editorial purposes. However, one can think of other contexts where this is an useful feature, namely in collaborative environments where multiple users contribute to a shared set of documents.

We produce summaries for a given period using a very simple approach based on the terms inserted between revisions. We only consider the oldest (D_{old}) and the newest version (D_{new}) of the document being analyzed within a defined period (e.g. month, day, or other). After extracting all paragraphs and removing all HTML tags, we prepare two sets of terms ($\{T_{old}\}$ and $\{T_{new}\}$), including each term's frequency. Terms are single words (unigrams) and groups of two consecutive words (bigrams) occurrences.

A set representing all inserted terms ($\{C_{old,new}\}$) is built by subtracting the old terms frequency count from the new terms frequency count. Each term is then scored by their resulting term frequency within $\{C_{old,new}\}$. The final ranked list of terms can be presented to the final user or used as input to a sentence selection algorithm. Since we only consider the first and the last version of a given period (all intermediary revisions are ignored), the algorithm's performance does not depend on the number of revisions. The algorithm's complexity depends on the article's size and has an upper bound of $O(N \times log(N))$.

4. WIKICHANGES

We developed a web-based software system to produce update profiles for Wikipedia articles, named WikiChanges. Perl was the primary programming language used and the graphics are based on amCharts[3], a dynamic charting library in Flash. The overall architecture of the system is shown in Figure 2. Although the system supports the simultaneous plot of two update profiles, the figure illustrates the flow for a single query.

When a query is presented to the system, the first step is the association of the query with a single article. This is done by searching a local file containing the names of all articles in the English Wikipedia. Then, the system checks if the local cache already has a copy of the updates to the article. If not, a background process is started to access the Wikipedia API and download the revision history for the given article. Since this is usually a long process, the user is informed and redirected to a temporary web page. If, on the other hand, a local copy is found, the system only downloads the latest updates while the user waits for the web page to finish loading. Before redirecting the user to the final web page, aggregated update values by month and day are calcu-

[2] `http://en.wikipedia.org/w/api.php`

[3] `http://www.amcharts.com/`

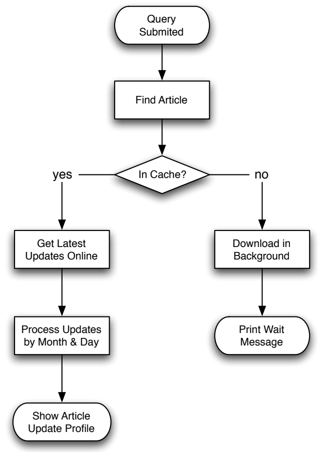

Query Submited
Find Article
In Cache?
yes
no
Get Latest Updates Online
Download in Background
Process Updates by Month & Day
Print Wait Message
Show Article Update Profile

Figure 2: Flow Chart for WikiChanges.

lated and written to the cache. Since WikiChanges is used in a multi-user environment, it was necessary to implement a simple concurrency control feature to avoid simultaneously downloading the same files.

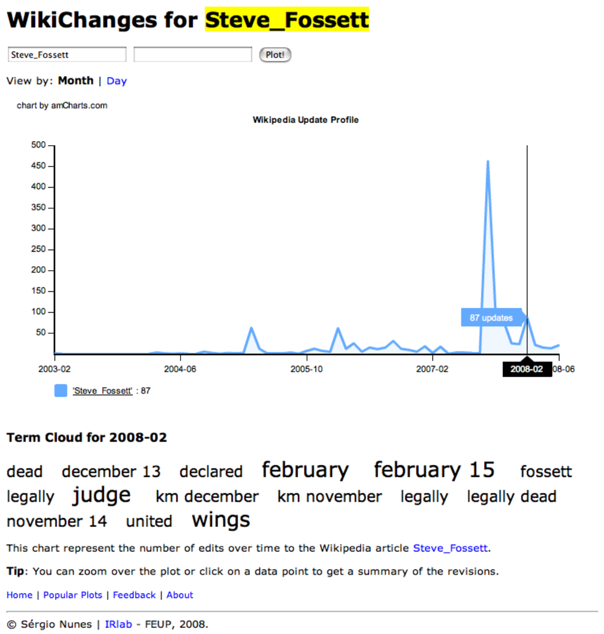
All data is stored using flat files. For each article, a file containing all Wikipedia updates is built as shown below in an excerpt. The first field is the revision ID as defined by Wikipedia and the second field is the time at which the revision occurred.

```
173900445 2007-11-26T15:22:08Z
173869943 2007-11-26T11:17:31Z
173869801 2007-11-26T11:16:02Z
173869695 2007-11-26T11:14:55Z
173868199 2007-11-26T11:01:35Z
...
```

The amCharts library, shown in use in Figure 3, is very flexible and has a number of useful built-in features. It's possible to navigate through the graph and have access to the individual values in each dataset as dynamic tooltips. The user is allowed to zoom in a portion of the data by selecting a specific range using the mouse. Also, it is possible to associate an URL with each single data point. We've used this feature to allow the user to view the automatic summary for a specific period. If the user clicks on a data point, he is redirected to a page where the month's summary is shown.

Figure 3 shows the Wikipedia article on the late USA adventurer *Steve Fossett* as viewed in WikiChanges. The summary shown for February 2008 is presented as a *tag cloud* in the bottom of the screen. This summary is automatically generated online using the algorithm described in the previous section. Due to the simplicity of the algorithm, this is feasible in real-time without caching the results. The size of each term is plotted in a logarithmic scale based on the term score.

In the example shown, the terms highlighted are related to the declaration of Steve Fossett's death by a judge on the 15th of February. It is important to note that, since terms are both unigrams and bigrams, some words appear multiple times within the tag cloud. For instance, the word *"february"* occurs isolated but also grouped with *"15"*. This tag cloud is a good example of an automatic summary produced by our algorithm. However, it is important to note that most


WikiChanges for Steve_Fossett

Steve_Fossett

View by: **Month** | Day

Term Cloud for 2008-02

dead december 13 declared february february 15 fossett legally judge km december km november legally legally dead november 14 united wings

This chart represent the number of edits over time to the Wikipedia article Steve_Fossett.

Tip: You can zoom over the plot or click on a data point to get a summary of the revisions.

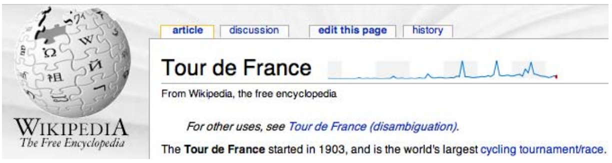
Home | Popular Plots | Feedback | About

© Sérgio Nunes | IRlab - FEUP, 2008.

Figure 3: WikiChanges system showing the revision history for the Wikipedia article on *Steve Fossett.*

tests conducted returned ambiguous or generic summaries. This is further discussed in the conclusions of the paper.

To increase the visibility of an article's update profile we developed a Gresemonkey extension. Greasemonkey[4] is a plug-in for the Firefox web browser that allows the user to "*customize the way webpages look and function*". Our Greasemonkey extension adds a sparklines to Wikipedia article pages as shown in Figure 4. Sparklines are "*small, high resolution graphics embedded in a context of words, numbers, images*" proposed by Tufte [4] to succinctly present data where it is discussed. We opted to show the sparklines next to the article's title for higher visibility and context. The sparkline was built using WikiChanges and the Google Chart API[5] for improved response time.



Tour de France

From Wikipedia, the free encyclopedia

For other uses, see Tour de France (disambiguation).

The **Tour de France** started in 1903, and is the world's largest cycling tournament/race.

Figure 4: WikiChanges sparkline embedded in Wikipedia's article *Tour de France.*

The interactive WikiChanges tool and the Greasemonkey script are available online at `http://sergionunes.com/p/wikichanges`.

5. RELATED WORK

[4] `http://www.greasespot.net/`

[5] `http://code.google.com/apis/chart`

In recent years, information visualization (*infoviz*) has gained popularity as a tool to explore the information contained in large datasets. Infoviz tools can be used to expose hidden patterns in large or complex datasets. Wikipedia holds a vast repository of data, most of which unavailable the casual user. In this section we briefly presents works where infoviz techniques have been applied to the Wikipedia.

Suh et al. [3] present WikiDashboard, a tool which "*aims to improve social transparency and accountability on Wikipedia articles*". The dashboard is embedded in live Wikipedia articles and combines information about the article's revision history and each contributor activity. Although similar to our work, WikiChanges differs in three main aspects. WikiDashboard is focused on recent activity, while WikiChanges shows the complete revision history for each article. Thus, in WikiChanges it's easier to spot recurring trends in articles as well as sporadic peaks. Also, we address a new challenge in the context of Wikipedia, namely *changes summarization*. We enunciate the problem and propose a simple approach based on the *term difference* between the start and end revision. Finally, our approach uses a richer graphical user interface containing several interactive features (e.g. zooming and side by side comparison).

WikiRage[6] is a "*site [that] lists the pages in Wikipedia which are receiving the most edits per unique editor over various periods of time*". WikiRage can be seen as a tool to identify current hot topics in Wikipedia. On the other hand, WikiChanges is an exploratory tool that is focused on single article analysis.

Wilkinson et al. [6] studied the editorial activity of high-quality Wikipedia articles and conclude that these articles "*are distinguished by a marked increase in number of edits, number of editors, and intensity of cooperative behavior, as compared to other articles of similar visibility and age*". Wikipedia featured articles were used to identify high-quality texts. On a macroscopic level, the authors found that edits to Wikipedia articles follow a stochastic process, where "*the number of new edits to a given article in a given period of time is a randomly varying percentage of the total number of previous edits*". Complementing these findings, we show that at the article level the editorial activity is far from random, being highly associated with the popularity of the topic.

Gawryjołek et al. [2] also identify an increased editorial activity in Wikipedia articles when the time of the year corresponds to the content mentioned in the text. They present JWikiVis, a desktop visualization software that "*helps to understand how collaborative documents are created and how they evolve over time*". Finally, in this same line of research, Viégas et al. [5] study collaboration patterns within wiki systems using a tool named *history flows*.

6. CONCLUSIONS

In this paper we present WikiChanges, a web-based tool for exploring Wikipedia article's revision history. We show that the update profile of Wikipedia articles can hold valuable information that is hard to capture using MediaWiki's standard interface. WikiChanges exposes this information and introduces several innovative features, most notably the summarization of changes for a given period, side by side comparison of update profiles, and embedded sparklines for the live Wikipedia. We think that the embedded sparkline feature is a useful interface improvement that could easily be incorporated in MediaWiki. Initial feedback from real users has been very positive.

Changes summarization is described as a task in the context of dynamic document collections. We propose a simple approach and present an example of a summary produced by WikiChanges. One limitation found while studying the construction of automatic summaries was the diversity of intentions of the users that contribute to an article. When an event draws attention to a given Wikipedia article, two types of positive contributions seem to be made: updates on the specific event and generic revisions to the whole topic. For instance, the recent death of Luciano Pavarotti resulted in more than 600 new revisions to the related Wikipedia article. However, most of these revisions were not related to this particular event, but instead were general updates to Pavarotti's profile. This is an important challenge for summarizing changes in highly active contexts.

In future work we intend to test new algorithms for changes summarization in dynamic collections. Also, we plan to evaluate the summarization results obtained using human assessors. The approach presented here only considers the first and last revision made to an article, ignoring the integral revision history. Further, we would like to test WikiChanges in other wiki contexts (e.g. documentation repositories, small team wikis, authenticated wikis). Particularly, we plan to investigate if most of our finding still hold in different contexts.

7. ACKNOWLEDGMENTS

Sérgio Nunes was financially supported by Fundação para a Ciência e a Tecnologia (FCT) and Fundo Social Europeu (FSE - III Quadro Comunitário de Apoio), under grant SFRH/BD/31043/2006.

8. REFERENCES

[1] Anonymous. Wikipedia - wikipedia, the free encyclopedia. http://en.wikipedia.org/wiki/Wikipedia [Visited 2008/05/05].

[2] J. Gawryjołek and P. Gawrysiak. The analysis and visualization of entries in wiki services. In *Advances in Intelligent Web Mastering*, pages 118–123. Springer Berlin / Heidelberg, 2007.

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

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

[5] F. B. Viégas, M. Wattenberg, and K. Dave. Studying cooperation and conflict between authors with history flow visualizations. In *CHI '04: Proceedings of the 2004 conference on Human factors in computing systems*, pages 575–582. ACM Press, 2004.

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

[6] http://www.wikirage.com/