Synchronicity - Automatically Rediscover Missing Web Pages in Real Time

Martin Klein
Computer Science Dept
Old Dominion University
Norfolk, VA, 23529
mklein@cs.odu.edu

Moustafa Emara
Computer Science Dept
Old Dominion University
Norfolk, VA, 23529
mal@cs.odu.edu

Michael L. Nelson
Computer Science Dept
Old Dominion University
Norfolk, VA, 23529
mln@cs.odu.edu

ABSTRACT

Missing web pages (pages that return the 404 "Page Not Found" error) are part of the browsing experience. The manual use of search engines to rediscover such pages can be frustrating and unsuccessful. We introduce **Synchronicity**, a Mozilla Firefox add-on that supports the Internet user in (re-)discovering missing web pages in real time.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Measurement, Performance, Design, Algorithms

Keywords

Synchronicity, Web Page Discovery, Memento, Preservation

1. INTRODUCTION

"404 Page Not Found" is a response frequently experienced when requesting a document on the web. Re-visiting bookmarks created some time ago or simply following a link from a poorly maintained web page may resolve in this error. Despite guidance for how to create "Cool URIs" that do not change [1] there are many reasons why URIs or even entire websites break [4].

It is commonplace for content to "move" to different URIs over time. The web page shown in Figure 1(a) for example returns a 404 error today. It used to provide information about the libertarian candidate William (Bill) E. Morris running for office in the US House of Representatives in 2004. Internet users are left with search engines to fulfill their information need. That however can be a frustrating endeavor since for example the query *bill morris* returns more than eight million results in a Google search.

In this paper we introduce **Synchronicity**, a Mozilla Firefox add-on that offers six options to support the Internet user in (re-)discovering missing web pages. Synchronicity utilizes *Memento*[1], a framework for time based access of web resources. When the user encounters a 404 response Synchronicity obtains a Memento TimeMap, a representation of an logical aggregation of all available Mementos (copies of web resources) of the missing URI. Mementos are aggregated from various sources such as search engine caches and web archives with the Internet Archive probably being the most famous contributor. With the help of Memento Synchronicity can discover the missing page at its new URI and provides options to retrieve good enough replacement pages in real time. This system is based on the intuition that information on the web is rarely completely lost, it is - since Mementos exist - just missing.

JCDL'11, June 13–17, 2011, Ottawa, Canada.

2. SYNCHRONICITY

Synchronicity is an extension to the Mozilla Firefox web browser. As mentioned earlier it triggers whenever the user encounters a 404 response from a web server and offers the following six methods to rediscover a missing page:

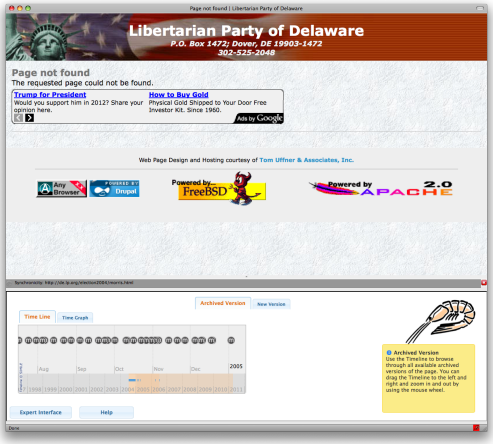
Option 1: The software retrieves a Memento TimeMap containing references to all Mementos. Synchronicity visualizes all available Mementos on a timeline from which the user can chose to display any particular Memento. If her information need is satisfied nothing else needs to be done.

Options 2 and 3: If that is not the case Synchronicity extracts the titles of the obtained Mementos and generates lexical signatures from their textual content. The system can use either as query strings against search engines. We have shown previously [3] that both web pages' titles and their lexical signatures perform very well as search engine queries to re-discover the page. Again the user decides whether the search results are sufficient. If no Mementos are available the software applies more complex methods to acquire a notion of the "aboutness" of the missing page.

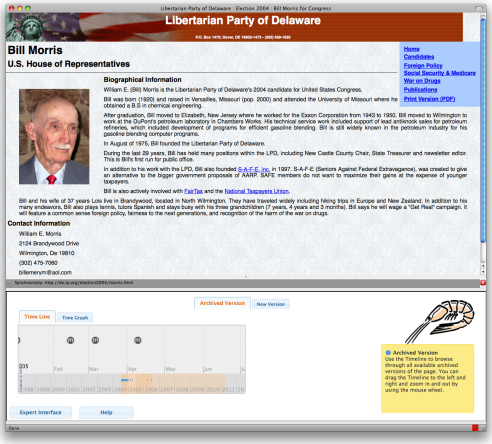
Option 4: It queries the social bookmarking site *Delicious* to obtain tags that users have used to annotate the (now missing) page. Tags have been shown to be useful for search [2] and so Synchronicity offers that method as well.

Option 5: The most complex option is to use the content of pages linking to the missing page as demonstrated in [5]. The assumption is that the aggregate of those pages is likely to be "about" the same topic. Synchronicity queries search engines to obtain backlinks (pages linking to the missing page) and generates a lexical signature from this link neighborhood. This signature also serves as a search engine query.

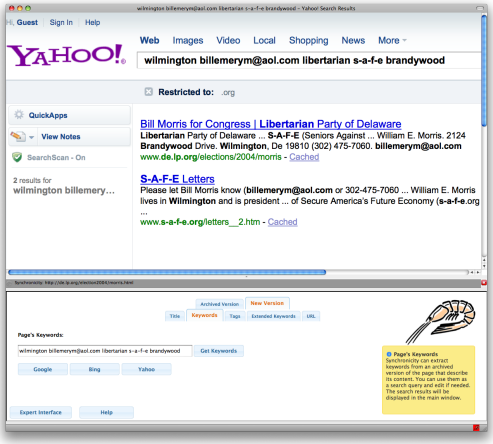
[1] `http://www.mementoweb.org/`



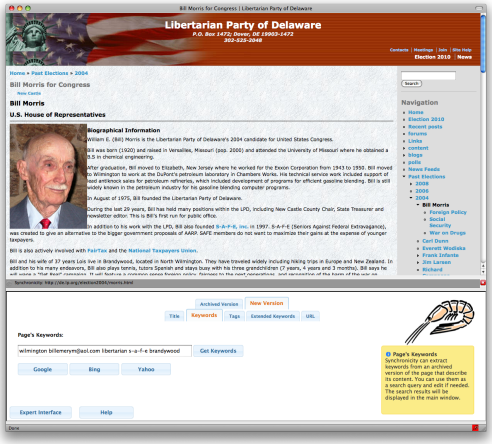
(a) Lost at:
`www.de.lp.org/election2004/morris.html`



(b) Memento from April 2005



(c) Query with *wilmington billemerym@aol.com libertarian s-a-f-e brandywood*



(d) Discovered Page at its New URI
`www.de.lp.org/elections/2004/morris`

Figure 1: Rediscovery of Bill Morris' Website

Option 6: Synchronicity at any time offers the option to modify the missing URI and retry. Our intuition is that shortening a long URI may help to at least find a new starting point to browse for the desired resource.

An important benefit of Synchronicity is that it works while the user is browsing providing results in real time. Options 4, 5 and 6 can be applied when no Mementos are available. Figure 1 shows Synchronicity applied to our example of Bill Morris' missing page. It activates on a server's 404 response 1(a) and displays the "zoomable" timeline with the dates of all Mementos. The latest Memento is displayed in 1(b) and the query with the generated lexical signature against Yahoo! in 1(c). Depending on the users information need option 1 may have already been sufficient. However, the top result of option 3 is the new URI (1(d)) of the page dereferencing the same content as the missing URI.

3. ACKNOWLEDGMENT

This work is supported in part by the Library of Congress.

4. REFERENCES

[1] T. Berners-Lee. Cool URIs don't change, 1998. `http://www.w3.org/Provider/Style/URI.html`.

[2] K. Bischoff, C. Firan, W. Nejdl, and R. Paiu. Can All Tags Be Used for Search? In *Proceedings of CIKM '08*, pages 193–202, 2008.

[3] M. Klein and M. L. Nelson. Evaluating Methods to Rediscover Missing Web Pages from the Web Infrastructure. In *Proceedings of JCDL '10*, pages 59–68, 2010.

[4] C. C. Marshall, F. McCown, and M. L. Nelson. Evaluating Personal Archiving Strategies for Internet-based Information. In *Proceedings of IS&T Archiving '07*, pages 48–52, 2007.

[5] J. Ware, M. Klein, and M. L. Nelson. An Evaluation of Link Neighborhood Lexical Signatures to Rediscover Missing Web Pages. Technical Report arXiv:1102.0930v1, Old Dominion University, 2011.