2014 2nd International Symposium on Computational and Business Intelligence

Data Reconstruction of Abandoned Websites

Iztok Fister Jr.[a], Iztok Fister[a], Simon Fong[b], Yan Zhuang[b]
[a] University of Maribor, Faculty of Electrical Engineering and Computer Science, Smetanova 17, 2000 Maribor, Slovenia
Email: iztok.fister2@uni-mb.si
[b] University of Macau, Faculty of Science and Technology, Macau, Av. Padre Tomas Pereira, Taipa, Macau
Email: ccfong@umac.mo

***Abstract*— Nowadays, the Internet offers data to anyone at any time. Websites on the Internet have been warehousing data for many years ago, i.e., for 10 years and more. In the meantime, many websites have became obsolete. This means they no longer have owner because of either they have no-one to maintain them or they have become unavailable for indexing by spiders that retrieves information about documents to be referenced. As a result, these websites are lost for accessing from Internet browsers and are therefore, referred to as abandoned websites. This paper focuses on the problem of how to identify the abandoned websites and how to preserve and reconstruct the data they hold. We have mainly concentrated on abandoned sport websites that, in general, contains very important data about the results achieved at various sporting competitions in the past. The proposed solution consist of four steps: an analysis of the abandoned servers that held these websites, identifying the structure of the abandoned web page sets, web scrapping, and preserving and visualizing these page sets. In order to test prototype solution, some steps were applied inorder to reconstruct and preserve the data on the abandoned web servers for tracking the results on running. Additionally, opportunities and challenges of applying data mining techniques on reconstructed website are listed.**

***Index Terms*— Internet, reconstruction, abandoned websites, web scrapping, data mining**

I. Introduction

Nowadays, the Internet is the most important source of information. People all over the world can obtain information everywhere at anywhere using the Internet browsers. Moreover, these browsers are still supported on all kinds of mobile devices [16]. People use the Internet in order to read news, to search for the upcoming social events, to arrange meeting with friends, to share photos, to watch television, to use for studying purposes, etc. Currently, the Internet contains millions of data terabytes and still increasing dramatically [13]. Data is contained either in data bases running on servers or physically on HTML sites that are available for direct presentations on web browsers. Recently, many web spiders have been arisen that traverses the web page's hypertext structure by retrieving a document and recursively retrieving all documents that have been referenced [1]. This data are saved into the databases of particular search engines (e.g., Google, AltaVista, etc.) and with their help users can browse for information. The search engine redirects the user to the homepage after finding the information in this base.

The majority of the newer websites are successfully maintained by spiders and thus well used by Internet users. On the other hand, the websites created many years ago (e.g., 10-15 years ago) are no available. When such websites are reorganized or migrated to a new platform some links in the page set can be lost in general. As a result, such websites cannot be accessed from the Internet and therefore, acts as a broken links. Although information in these websites exists it is not available to Internet users. In line with this, these websites are also inaccessible for spiders. Additionally, the spiders cannot index the websites even when they are not available because of network's and server's errors, or even when they are excluded from the search by the website designer. In our case, these websites are also called abandoned websites, while the server in which they host is an abandoned server. However, these abandoned websites can have held important data in the past (e.g., news, blogs, reports, sport results, events, etc.).

This paper examines the problem of recovering the abandoned websites. In line with this, a novel method is proposed that consists of four steps:

- the analysis of an abandoned server,
- identifying the structure of an abandoned web page set,
- web scrapping,
- data reconstructing and visualizing the recovered web page set.

The proposed method was tested by reconstructing and visualizing a recovered web page set hosted on an abandoned server for tracking the results of running competitions. These are especially important to for competitors who would like to track their progresses throughout the past years. The tests showed that data on abandoned websites can be successfully recovered using this method.

The structure of the remainder of this paper is as follows. In Section 2, the problem of abandoned websites is discussed as they arise on abandoned servers hosting the results achieved during running competitions. Section 3 proposes a method for recovering the abandoned websites. In Section 4, a case study is performed for recovering the results achieved during running competition. The paper concludes by summarizing the obtained results and outlining the directions for the further development.

978-1-4799-7552-5/14 $31.00 © 2014 IEEE
DOI 10.1109/ISCBI.2014.22

II. Abandoned websites

A lot of websites on the internet were created in the past. Unfortunately, many of these websites have been abandoned today due to the lack of maintenance. In fact, these websites were maintained locally. During the time, a lot of the people responsible for maintenance were assigned to another jobs. In general, these websites were not documented or the documentation was not sufficient. Therefore, any upgrading of the websites urgently leads to failures that were manifested as abandoned websites.

Typically, an abandoned website emerges when the following conditions are fulfilled:

- the website was created in the past and has not been updated over for several years,
- there are no more maintainers for this website,
- website were run on older technology that is no longer supported by the web server,
- developers have had no motivation to maintain the website because there was no payment for this job,
- data on the websites were not of interest for a wide-range of users.

As a result, the abandoned websites had no owner, who would have an interest in maintaining and even upgrading them. This means, that these websites are dead for the Internet users. Fortunately, many of these abandoned websites contain such important content that cannot be lost for the majority of web users. An International Internet Preservation Consortium was created in order to cope with this problem in general [15].

In the remainder of this paper, our method for data reconstruction of abandoned websites is presented in detail.

III. Data reconstruction of abandoned websites

The proposed method for data reconstruction of abandoned websites consists of four steps, as follows:

- the analysis of an abandoned server,
- identifying the structure of the web page set,
- web scrapping,
- preserving and visualizing the reconstructed data.

The proposed steps are also illustrated graphically in Fig. 1. In the first step, the file system on disk of the abandoned server is analyzed. Then, the structure of the web page set on disk is reconstructed. In this step, all files containing the active websites are tagged. All untagged files represents abandoned websites and are subject of web scrappers that can extract useful content (data) from the website. Finally, the extracted data are preserved into a database and displayed in the appropriate web browsers using a new form. Thus, two benefits are created. Firstly, data are archived for the future usage and secondly, they are represented a new using the state-of-the-art technology.

A. The analysis of an abandoned server

Analyzing an abandoned server is the initial step of data reconstruction that is crucial because this step depends on the destiny of the whole process. If data cannot be downloaded to the local server the process cannot be started.

Primarily, this analysis tackles the abandoned server platform and disk file system where the abandoned websites are hosted. Today, there are many platforms for hosting the websites. Obviously, on the platform it depends which system tools are used for accessing data within an appropriate file system. Typically, the *nmap* utility is used for this analysis [3] because it is independent of the platform on which it is used.

After this analysis, the tools are selected for downloading data onto the local disk. Here, the *wget* utility can be used in order that all the data are retrieved from the abandoned server.

B. Identifying the structure of web page set

Identifying the structure of the web page set starts by identifying home page. Typically, this page is found under the name *index* with extensions like *.html, .htm, .php, .asp, etc*. However, the name of this file can be distinguished from *index*. This is because this filename usually depends on the context in which it is created. Therefore, this identification step is performed manually.

Normally, the web page set forms a kind of tree, where pages are linked between each other using the HTML command $\langle A\ href = "Pagex.html" \rangle$ [4]. A script file as illustrated in Algorithm 1 is used in order to reconstruct the structure of the whole web page set. This script builds the document tree by visiting each page referenced in the visited document by *href* command (Fig. 2).

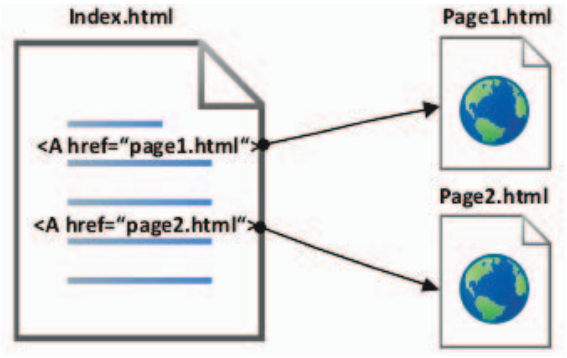

Figure 2. Structure of web page set

As can be seen from Algorithm 1, a building of the document tree starts with the *index* page and by recursively visiting all the referenced documents until a document without any link to reference document is detected (also a leaf document) (line 1). All the documents in this tree are placed into a set of tagged documents (lines 2-5).

An abandoned document set is determined as a difference between the tagged and all disk files. In this case, the list of files in the home directory and corresponding

Figure 1. Data reconstruction of abandoned websites

Algorithm 1 Script for selecting the abandoned websites

Input: Home page *index*, home page directory *dir*.
Output: Set of abandoned websites *AbandonSet*.

1: *PageSet* = Build_web_page_set(*index*);
2: $TagSet = \emptyset$;
3: **for all** $file \in PageSet$ **do**
4: $TagSet = TagSet \cup file$;
5: **end for**{Create the document tree}
6: *DirSet* = List_files_on_disk(dir);
7: $AbandonSet = \emptyset$;
8: **for all** $file \in TagSet$ **do**
9: **if not** $file \in PageSet$ **then**
10: $AbandonSet = AbandonSet \cup file$;
11: **end if**
12: **end for**{Create the abandoned document set}
13: **return** Select_websites_only($AbandonSet$);

sub-directories *DirSet* is obtained (line 6). Then, each file in this set is visited. If the visited file is not tagged it is placed into a set of abandoned documents *AbandonSet* (lines 7-12). However, because not all files on the disk represent valid websites only files with extensions like *.html, .htm, .php, .asp, etc.* are declared as abandoned websites (line 13).

C. Web scrapping

Web scraping is a computer software technique of extracting information from websites [10]. It refers to an application that processes the HTML of a web page in order to extract data for further manipulation, e.g., converting the web page to another format (i.e. HTML to WML). Web scraping scripts and applications simulates a person viewing a website with a browser. Using these scripts the person connected to a scraped website and view the same as in the original browser, however in reformatted form.

Today, many web scrapers exist on the market which are written in different programming languages. Python [5], [6] is an especially useful programming language for the developing the web scrapers. Therefore, Google Inc. that invests considerable effort in this area developed a *webscraping* Python library for web scraping [8]. On basis of this library, Penman et al. [2] developed a web scraper with the name SideScraper. In addition, Scrapy [7] represents a fast high-level screen scraping and web crawling framework, used to crawl websites and extract structured data from their pages. It can be used for a wide range of purposes, from data mining to monitoring and automated testing.

In our case, the web scraping is used in order to process the abandoned web pages in older HTML format and scrap data into XML format. Therefore, a simple web scraper using Python library for web scraping that on web pages searches for a specific patterns denoting the HTML orders for creating HTML tables. As is well known, HTML tables are formatted using orders like $\langle table \rangle$, $\langle tr \rangle$, $\langle td \rangle$, etc. The web scraper then converts these tables to the XML tables [11].

D. Preserving and visualizing

Our simple web scrapper extract data from the older abandoned HTML websites in XML formatted files. In order to preserve the extracted data on the one hand and to ensure the proper reliability on the other hand, the XML files need to be imported into relational database (e.g., MySQL) using Python script. This script is relatively simple and therefore not presented in the paper in detail.

Extracting data into XML formated files and putting these into a relational database only serves for preserving the older websites. How these data are presented to the users in web browsers depends on the visualizing step. During this step, some designer tasks need to be performed, in which the formats of the displayed reconstructed web pages are adjusted to the new state-of-the-art-technology.

Finally, these separate web pages need to be linked together into a single web page set and thus enabling users to browse the reconstructed websites simply and intuitively.

IV. Case study: Data Reconstruction of abandoned websites for tracking the running competitions

This case study was tackled the data reconstruction of websites for tracking those running competitions that were held under the auspices of the Slovenian Olympic Committee. Typically, these competitions were conducted

locally and therefore suffered from a lack of central control, which would be provided by a central backup for the achieved results of competitors. Currently, this website is alive but content was not updated for some years. A lot of results are also not available to see via browser. Therefore, we might say that this website is abandoned [9]. However, we can still save the content and store for future generations, before server break down and totally dissapear from network.

A goal of our experiments was to show that the abandoned websites could be successfully reconstructed. In line with this, we were focused on websites for tracking a regional cup in running that was held within the region of Prekmurje (Slovenia). The regional cup consisted of 10-20 runs during the whole year. This means that during each year the 10-20 abandoned websites could be produced.

The contents of these websites became more important for the competitors every day. Many competitors would like to compare their present results with their performances as achieved some years ago. This is especially important because in the past many of today active competitors did not have any chance to track their achieved results on the Internet. Today, the Internet for these competitors presents a warehouse, where all past achieved results can be obtained for any competitors in any competition. Therefore, it is so crucial to reconstruct the abandoned websites containing past results.

In order to prove the proposed method of data reconstruction, the known abandoned server hosting the results for the running regional cup was taken into account. The server analysis showed that it ran on a Windows XP Server SP2. The home directory consisted of files as illustrated in Fig. 3.

```
# ls .
 index.html
 2002.css
 default.asp.html
 calender.asp.html
 news.asp@n=11.html
 partizipation.asp.html
 photogalery.asp.html
 photogalery_display.asp@id=1001&f=0.html

 photogalery_display.asp@id=1001&f=6308.html
 run_results.asp@id=1001&f=0.html

 run_results.asp@id=1001&f=1379.html
```

Figure 3. An analysis of abandoned server

From a list of files in home directory, the index file could easily be determined because it held the standard name *index.htm*. The script for identifying the structure of the web page set created the document tree structure as presented in Fig. 4.

It can be seen from Fig. 4 that the maximal depth of the document tree is one. This means, the structure is very simple because all pages in the set are directly referenced by an index page.

```
-index.html
  -run_results.asp@id=1360.html
  -run_results.asp@id=1361.html
  -run_results.asp@id=1362.html
...
  -run_results.asp@id=1379.html
```

Figure 4. Identifying the structure of web page set

```
# ls Reconstruction-Abandon/Html/
 index.html
 2002.css
 default.asp.html
 calender.asp.html
 news.asp@n=11.html
 partizipation.asp.html
 photogalery.asp.html
 photogalery_display.asp@id=1001&f=0.html

 photogalery_display.asp@id=1001&f=6308.html
 run_results.asp@id=1001&f=0.html

 run_results.asp@id=1001&f=1359.html
#
```

Figure 5. Create the abandoned document set

The abandoned document set can be seen in Fig. 5. Note that this document set consists of many files that do not represent websites, e.g., cascade style sheet *.css*, etc. These files are, however, not members of the abandoned website set.

The web scraper took each website in HTML format from the abandoned website set and extracted their content into XML files. Thus, it focused on the structure of HTML table (Algorithm 2). The content of HTML table was extracted into an XML tagged structure as presented in Algorithm 3.

Algorithm 2 HTML original abandoned website

```
1: ⟨table width=100⟩
2: ⟨tr⟩ ⟨td⟩ ⟨br⟩ ⟨/td⟩ ⟨/tr⟩
3: ⟨tr⟩ ⟨td⟩
4: ⟨pre⟩
5: ⟨b⟩Boys and girls under 2002⟨/b⟩
6: Boys
7: Place Surname    Name     Club        Age  Time
8: 1     Župančič   Klemen   TD Bovec    2002 1,30
9: ...
10: 5    Mavrič     Martin   TD Bovec    2003 1,45
11: ⟨/pre⟩
12: ⟨/td⟩ ⟨/tr⟩
13: ⟨tr⟩ ⟨td⟩ ⟨br⟩ ⟨/td⟩ ⟨/tr⟩
14: ⟨/table⟩
```

Finally, the reconstructed website was introduced to the user in a new format as presented in Fig. 6. As can be seen in this figure, the data reconstruction was successfully completed. Furthermore, the reconstructed data are also available to the other application because these are still held in a database.

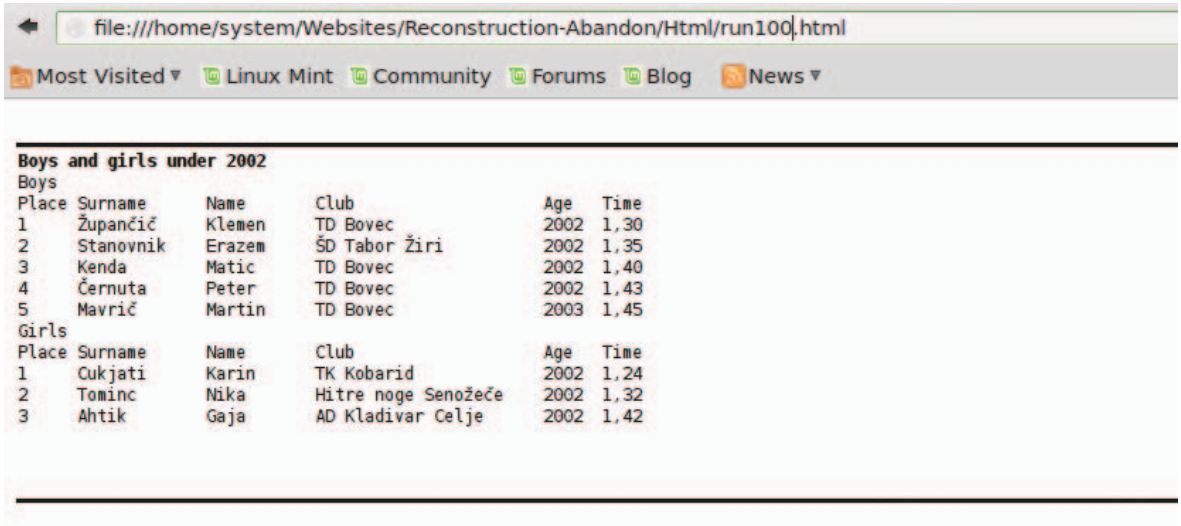

Figure 6. Extracted data in web browser

Algorithm 3 XML reconstructed abandoned website

```
-⟨competition⟩
  ⟨category⟩Boys and girls under 2002⟨/category⟩
  -⟨competitor⟩
    ⟨place⟩1⟨/place⟩
    ⟨surname⟩Župančič⟨/surname⟩
    ⟨name⟩Klemen⟨/name⟩
    ⟨club⟩TD Bovec⟨/club⟩
    ⟨age⟩2002⟨/age⟩
    ⟨time⟩1.30⟨/time⟩
    ⟨sex⟩male⟨/sex⟩
  ⟨/competitor⟩
  +⟨competitor⟩ ⟨/competitor⟩
  ...
  +⟨competitor⟩ ⟨/competitor⟩
⟨/competition⟩
```

A. Discussion and future challenges

The results from the proposed method for the data reconstruction of abandoned websites preserved the older, usually yet forgotten, websites on a server, extracted their content and displayed it in a new form. Although the performed case study was relatively simple, this method showed on all the problems with which someone who wishes to reconstruct data from older websites is confronted with. In this method the web scraper should obviously be improved in order to reconstruction of more complex websites would be possible. On the other hand, many of these web scrapers are already made and could be used in the proposed method. With this in mind, this method can be seen as an universal tool for the data reconstruction of abandoned websites.

B. Opportunities and challenges using data mining

When abandoned website is reconstructed, we can also apply some data mining techniques [18] in order to extract a 'hidden' knowledge from data. Especially, sport results are a good data for applying data mining, because a lot of sport scientists, sport analysts, sociologists compare current performances of athletes with performances in the past. They try to predict performance and trends for the future. On the other hand, from historical view it is also good to know training habits in the past and compare with current habits of athletes. Therefore, data mining is a nice tool for extraction of very important data from abandoned websites.

V. Conclusion

The problem of abandoning websites is still alive and therefore, many historical data are seen as being lost. In this study, we presented a novel solution for reconstructing data from abandoned websites. In line with this, we successfully analyzed an abandoned server, identified abandoned websites, extracted their useful content and preserved it in a database. Data in this database are then available for new visualization and to another applications as well. A case study on an abandoned server hosting the results of a running regional cup showed that this method could be a universal tool for the data reconstructions of abandoned websites.

In the future, this method would be more elaborated on, especially by including the new web scrapers. Moreover, exploring new ways of preserving old and abandoned websites will be explored in the future. One of the very promising possibility is to control parsing process of web scrapers with domain specific languages [17].

Acknowledgement

The authors are thankful for the financial support from the research grant of Grant no. MYRG152(Y3-L2)-FST11-ZY, offered by the University of Macau, RDAO.

References

[1] A.G. Lourenço, O.O. Belo, Catching web crawlers in the act, Proceedings of the 6th international conference on Web engineering, ICWE '06, Palo Alto, California, USA, pp. 265–272, 2006.

[2] R.B. Penman, T. Baldwin, D. Martinez, Web scraping made simple with Sitescraper, http://sitescraper.googlecode.com/, Accessed 2 March 2014.
[3] M. Wolfgang, Host discovering with nmap, http://bandwidthco.com/whitepapers/netforensics/recon/nmap/Host%20Discovery%20With%20nmap.pdf, Accessed 5 March 2014.
[4] J. Duckett, HTML and CSS: Design and Build Websites, Willey & Sons, NY, 2011.
[5] W. McKinney, Python for Data Analysis, O'Reilly Media, 2012.
[6] D.M. Beazley, Python Essential Reference (4th Edition), Addison-Wesley Professional, 2009.
[7] Scrapy, http://scrapy.org/, Accessed 6 March 2014.
[8] Webscraping, http://code.google.com/p/webscraping/, Accessed 3 March 2014.
[9] Slovenija tece, http://www.slovenijatece.si/, Accessed 20 February 2014.
[10] R. Song, Joint Optimization of Wrapper Generation and Template Detection, The 13th International Conference on Knowledge Discovery and Data Mining, 2007.
[11] R. Kuramdasu, http://www.codeproject.com/Tips/512500/Converting-XML-string-to-HTML-Table, Accessed 5 March 2014.
[12] P.J. Lynch, S. Horton, Web style guide: Basic design principles for creating Web sites, Yale University Press, 2008.
[13] Ao Ma, Yang Yin, Wenwu Na, Xiaoxuan Meng, Qingzhong Bu, and Lu Xu. Scrubbing in storage virtualization platform for long-term backup application. In *Availability, Reliability and Security, 2009. ARES'09. International Conference on*, pages 441–447. IEEE, 2009.
[14] L.E. Ullman, PHP and MySQL for dynamic web sites, Peachpit Press, 2003.
[15] International Internet Preservation Consortium, http://netpreserve.org//, Accessed 5 March 2014.
[16] I. Fister, D. Fister, S. Fong, I.Jr. Fister, Widespread Mobile Devices in Applications for Real-time Drafting Detection in Triathlons, Journal of Emerging Technologies in Web Intelligence, 5(3):310–321, 2013.
[17] I.Jr. Fister, I. Fister, M. Mernik, J. Brest, Design and implementation of domain-specific language Easytime, Computer Languages, Systems & Structures, Elsevier, 37(4):151–167, 2011.
[18] I. Fister Jr., D. Fister, I. Fister, S. Fong. "Data mining in sporting activities created by sports trackers." Computational and Business Intelligence (ISCBI), 2013 International Symposium on. IEEE, 2013.