International Institution for Technological Research and Development
Volume 1, Issue 4, 2016

Adware Removal using Real-Time Web Service Fault Injectioning.

Kavan Dave[1], Bhavesh Tanawala[2] and Hemant Vasava[3]

[1] Department of Computer Engineering, B.V.M. Engineering Collage,
Vallabh Vidhyanagar, Anand, India
kavandave92@yahoo.com

[2] Department of Computer Engineering, B.V.M. Engineering Collage,
Vallabh Vidhyanagar, Anand, India.
bhavesh.tanawala@bvmengineering.ac.in

[3] Department of Computer Engineering, B.V.M. Engineering Collage,
Vallabh Vidhyanagar, Anand, India
hdvasava@bvmengineering.ac.in

Abstract

Internet is a huge mine and warehouse for information storage and extraction. Adware on browsers are the most irritating things which may redirect user or install software secretly and cause harm to user's confidentiality and systems. So the removal of malicious adware is an important perspective. This research proposes a system to identify adware from the current web-page on the browser by comparing its HTML code with the actual code using the concept of HTML parser and web service fault injectioning.

Keywords: *JSoup; HTML parser; Web service; Adware; Fault injection..*

1. Introduction

In the era of technology internet is working as sorcerer which provides knowledge of everything on your fingertips. With the evolution and awareness of internet across the world advertisement popping has also become very popular these days. Adwares are the software which provides publicity on websites legally. But there are many malicious adware that had affected confidentiality, integrity, consistency and durability of information over the internet. It provides harm to users as well as their system. It exhausts network bandwidth, waste resources and also may access user's personal data. So these kinds of malicious adwares are needed to be removed.

Some lethal consequences [6] because of adware give remarkable impact on industry. Zango was an official adware provider on Facebook. Inspite of having legal SLA (Service Level Agreement) between Zango and Facebook, Facebook users have faced some illegal advertisement called "secret crush" which tempted many users to use it and got their malicious software installed on their computers. This affected 4% of Facebook users and this resulted into disagreement between both companies. In 1988 [7] due to adware the economical loss of $6.1 billion was reported which was increased and reached up to $13.3 billion in 2006 and also increasing in current scenario. So the removal of adware has become an important task in software engineering in terms of after deployment support.

To identify the adware from the pages the extraction of web-page information is required. Here a system is proposed with the use of HTML parser to extract web information with HTML and the mechanism of real-time web service injectioning. If the user of the website would like to check whether his page is infected or not, he would be able to invoke the fault injected web service which will compare the code of current open page on browser with the actual code available on the server and will provide the notification to user as well as the web site admin if the Adwares are available on the current page.

2. Background Theory

2.1 WEB SERVICE

The phrase "Web service" narrate a standardized way of integrating web-based applications using XML (Extensible Markup Language), SOAP (Simple Object Access Protocol), WSDL (Web Service Description Language) and UDDI (Universal Description, Discovery and

Integration) open standard over standard internet. SOAP is used for exchanging information, WSDL for description and UDDI works as a registry for web services [8].

2.2 HTML

HTML stands for Hyper Text Markup Language. In late 1991 Tim Berners-Lee created HTML. Today HTML is the most preferred language for developing web pages. The latest version of HTML is "HTML 5" which was published in 2012. HTML adds "Markups" to Standard English language and "Hyper Text" refers to the link for connecting one web page to another [9].

2.3 HTML PARSER

HTML Parser is a Java library by which you can parse your HTML code in either linear or nested fashion. It is very fast and robust tool. Basically it used for any transformation or extraction of data, but it also facilitates user with additional features of filter, visitor, custom tag and support for JavaBeans. The most attractive feature of HTML Parser from developer's perspective is its simplicity in design [10].

2.4 JSOUP

For working with real –time HTML web pages a library is needed to work in Java language. Jsoup provides the developer with a Java library for this purpose. Jsoup provides a very suitable API for the manipulation and extraction of data using DOM (Document Object Model), CSS (Cascading Style Sheet) and jquery-like methodology. It implements the WHATWG HTML5 specification and parses HTML to DOM [11].

2.5 ADWARE

Adware is a common name of all kinds of software presented to user which contains advertisement rooted with the application. It is a legitimate option for consumer who doesn't want to pay for software. In today's technocrat era, many developers offer "sponsored" freeware to user. If the adware is legitimate the ads will disappear as soon as you stop running your software. If the adware is not legitimate and contains malicious things it is called Malware [12]

3. Literature Review

Mr. Jie Yang and Yuancheng Li [1] have stated a novel method to extract informative blocks from web pages and filter the advertisement which has nothing to do with the subject when people browse the web page. They used HTML Parser to construct DOM tree and CST. This method works with high accuracy but it is complex to develop.

D.S.Patil and N.A.Dhawas in [2] have proposed enhancing automating extraction of top-k to find particular information or accurate data from web. They have extracted top-k list from all available database which contain data in either structured or unstructured format. This method is relatively accurate because extract top-k list but the flip side is difficulty in dealing with non contiguous data record.

Mr. Zhang Xu and Dong Yan in [3] have proposed designing and implementing of web page information extracting model based on tags to make the search engine more efficient by improving web spider to get useful resources like links and pictures quickly. For that they have analyzed the web page properly using HTML Parser and Jsoup based on Java. This system enhances the recall and lowers the precision but because they have developed strict system flow it can't deal with some imprecise websites properly.

Mr. Lin Shan and Zhang Qun in [4] flexible approach for web information extraction based on HTML Parser said a flexible and high approach to web information extraction. HTML Parser is parsing library which has been used to transform or extract the web information. This approach can be successfully implemented for custom tags also but having less accuracy.

Mr. Fang and Mr. Lyu said in [5] educational resources metadata automatically extracted strategy study about the development of obtaining information for efficiently from internet using the principle of HTML Parser and Java related regular expression knowledge. It works significantly with structured function and labels but label resolving still requires further study

4. Methodology

From the literature survey we studied that the HTML Parser can be used for feature extraction and adware removal. So here we have proposed and algorithm for removal of adware from the browsers working page using real-time fault injection in web services.

Algorithm:

- If user is using a website having a mechanism of real-time fault injected web service to check whether the pages infected with adware or not.
- User will use that service and the code of currently opened page and the actual code of that page will be compared by the web service format.
- If both the code is not similar a notification message will be sent to website owner as well as the browser user.

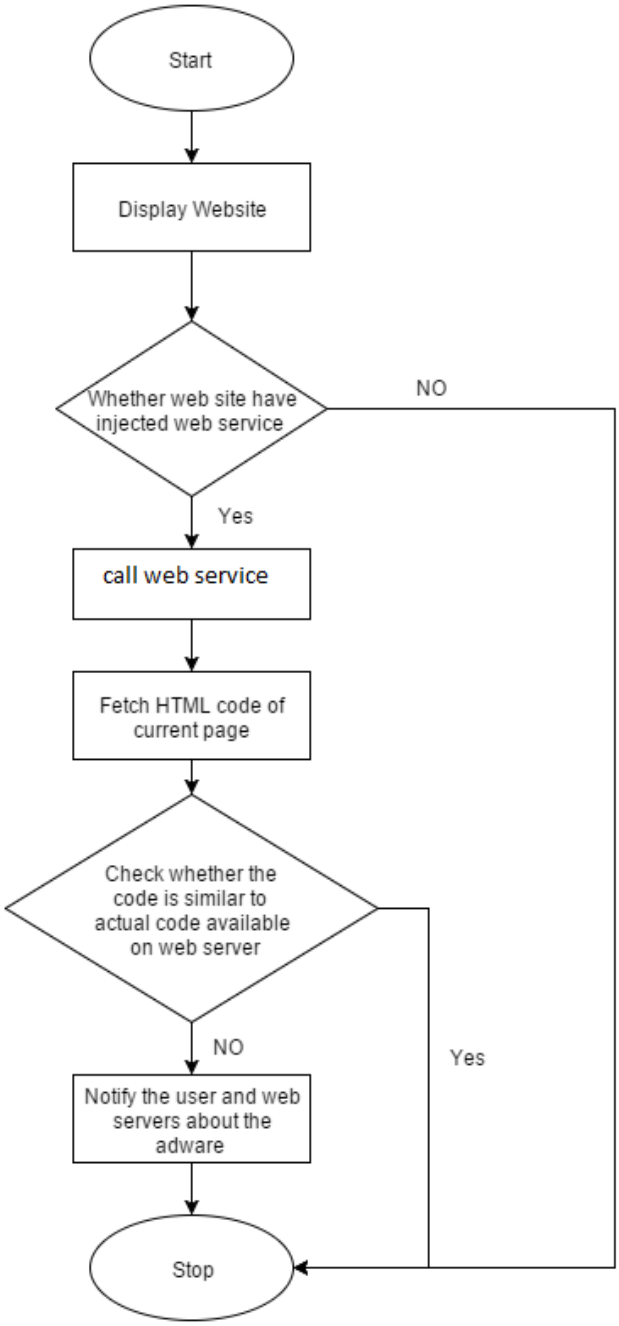

Figure 1: Flow for Adware Removal Mechanism.

4. Conclusions

We have proposed two web services to identify adware from the current web-page on the browser by comparing the current web page's HTML code with the actual code using the concept of HTML parser and web service

injectioning which works more efficiently on structured, semi-structured and unstructured data with low latency.

.

Still this method has a scope of improvement if any better substitute of HTML Parser can be applied.

Acknowledgments

It is a matter of great pleasure for me to get this opportunity of expressing my deep and sincere gratitude to those who has always helped me throughout my work. I heartily thank my guides Mr. Bhavesh A. Tanawala & Mr. H.D.Vasava for providing me their valuable guidance and constant support in my work. They guide me with their excellent knowledge in the area of "Software Engineering." I am also thankful to my friend Mr. Dhavalkumar H. Joshi for his constant cooperation, help and guidance throughout this journey. I would also like to thank all my colleagues who have helped me in some way or the other. I would like to thank my parents & family member for their love and support to build my moral during the work

References

[1] Yuancheng Li and Jie Yang, "A Novel Method to Extract Informative Blocks from Web Pages" IEEE 2009 International Joint Conference on Artifical Intelligence (IJCAI)-Beijing,China..

[2] Dipali S. Patil and Prof. N.A.Dhawas, "Enhancing Automatic Extraction of Top-k List from Web," IEEE 2014 International Conference for Converengence of Technology (ICCT) – Lonavala.

[3] Zhang Xu and Dong Yan, "Designing and implementation of the Web page Information Extraction based on tags", IEEE 2011 International Conference on Intelligence Science and Information Engineering,(ICISIE-2011) - Beijing, China.

[4] Lin Shan and Zhang qun, "Flexible approach for Web Information Extraction Based on HTML Parser," IEEE 2012 IEEE 7th International Conference on Computer Science and Education - Melbourne, Australia (2012.07.14 to 2012.07.17).

[5] Fang Yanfen and Liu Qingtang, "Educational resources metadata automatically extracted strategy study," IEEE 2008 International Symposium on Knowledge Acquisition and Modelling Applications (ISAKAM) - Wuhan, China.

[6] Types of Adware: Zango[Online]. Available:www.spanlaws.com/zango-adware.html

[7] Annual Worldwide Economic Damages from Malware Exceed $13 Billion [Online] Available: http://www.computereconomics.com/article.cfm?id=1225

[8] Web-Service[Online] Available:http://www.webopedia.com/TERM/W/Web_Services.html

[9] HTML[Online] Available:https://developer.mozilla.org/en-US/docs/Web/HTMLtp://www.webopedia.com/TERM/W/Web_Services.html

[10] HTML Parser[Online] Available: http://htmlparser.sourceforge.net/

[11] JSoup[Online] Available:ht http://jsoup.org/

[12] Adware[Online] Available: http://www.webopedia.com/TERM/A/adware.htmll

[13] S. Gupta, G.E.Kaiser, D.Neistadt, and P. Grimm, "Dom-based content extraction of html documents," The 12th international conference on World Wide Web, 2003, pp. 207-214, doi:10.1145/775152.775182.

[14] Lin Shian-Hua, and Ho Jan-Ming, "Discovering Informative Content Blocks from Web Documents," The eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002, pp. 588-593, doi: 10.1145/775047.775134.

[15] S. Debnath, P. Mitra, and C. L. Giles, "Automatic extraction of informative blocks from web pages," The 2005 ACM symposium on Applied computing, 2005, pp. 1722-1726, doi: 10.1145/1066677.1067065