Extending WebCred: Assessing the Credibility of Web Sites

Christopher Hopper
Computer Information Science
Benedict College
Columbia, SC 29206
bob12345629223@yahoo.com

Abstract

The evolution of the Internet has made the sharing of information essential to the intellectual growth of human society. Along with the ability to exchange information relatively quickly brings the question of credibility. The credibility of the authors, the quality of the information, and the validity of sources are some of the legitimate user concerns. An automatic assessment of credibility can offer some answers to questions about the credibility of information found on the Web. Through the use of an automated system the credibility of websites can be gauged based on a number of criteria. A program called WebCred has been implemented for this purpose. WebCred rates the credibility of URLs for the medical domain. This program rates the URLs based on the following factors: the number of page errors, credibility of the links, and the number of credible ads. This paper focuses on implementing WebCred for several domains and to provide end-users several levels of security from malicious HTML.

1. Introduction

Credibility of information is a complex issue due to the fact that it is largely based on the individual. For example, a professional who is seeking information on a specific topic may have biases that influence their willingness to accept the credibility of information submitted by a rival. Another example is a website that could be considered a credible source and yet contain inaccurate information. There are always credibility issues when searching for useful information. When doing research more than one reference is generally required to be considered as credible. The majority people may consider some information that comes from the government to be dependable, but there are always others who distrust the government and any information that is put out they consider to be untrustworthy.

A system for automated assessment of credibility can provide confidence in searching the Web. Users want to make sure the information they find is accurate and up-to-date. A tool like WebCred can shorten their time searching the Web by eliminating websites that it rates as not reliable. The program itself can be dependable if it does this task well and it safe to use. Safety in web searching is also vital to users. Some websites contain malicious links and data. The possibility of providing the detection of malware will make the program even more of an asset to use. Section 2 of this paper will discuss related work in this area. Section 3 will describe the current implementation of WebCred along with added features. The evaluation of the program's performance will be discussed in section 4. Finally, section 5 contains the conclusions of the research and a discussion of future work.

2. Related Work

Research that is related to WebCred involves two different sources. The first source is concerned with teaching Web evaluation for K-12 schools. It is primarily concerned with the following concepts:

- Why Evaluate Web Information
- Methods of Evaluation
- Teaching Web Evaluation

Here are some of the tools designed for educators teaching web evaluation for K-12: Kathy Schrock's Guide for Educators, Alexander and Tate's checklist, and the CARS checklist (Brown, et al., 2002). I searched all three of these sites and found some good criteria for web credibility. Kathy Schrock's site is a series of surveys for elementary to middle schools students. Alexander and Tate's checklist is really a list of practice tips for evaluating Web pages. They include determining the purpose/type of the site in the following categories: entertainment, business, information, news, advocacy, or personal presentation (Rutherford, 2004). The CARS website is an acronym for Credibility, Accuracy, Reasonableness, and Support (CARS, 2002). These tools are all fine criteria for evaluating Web Credibility but they are not trying to implement an automated system that supports their theories.

The second source is a Web credibility project for Stanford University. The group's research falls within four broad areas: evaluation strategies, design, individual factors, and contextual factors (Stanford, 2005). They will also attempt to (1) develop new experimental methodologies useful for investigating Web credibility, (2) synthesize existing research in the Web credibility literature, and (3) advance theory in the field (Stanford, 2005). This research will provide a great deal of information on how to perform credibility research on the Web. This is similar to the program, because it is for specialized users. The users of the WebCred program will often know the domain they want to pursue. The users of the Stanford research method are searching for specified information. The scientific approach to Web credibility will cut down on lost time spent on useless websites or useless information found on credible websites. The creation of a system would be a welcome addition to Stanford's research. All of the techniques that are being considered for human effort to perform could be applied to a system such as WebCred.

3.1 Implementation

The current program WebCred was written in Microsoft Visual C++ .NET by a former USC REU 2004 intern named Marcus Wassmer. The program is really part of a suite that encompasses three tools to achieve its goal. The first tool is an online validator program. This program is available on the W3C website (see references). The W3C Markup Validation Service program version 0.67 takes the URL entered by the user and finds the page errors with a SGML parser. The types of errors that are found are most often within the HTML syntax.
Figure 1

1. *Line 138, column 116*: **cannot generate system identifier for general entity "z"**

2. `...d.com/pages/dir/toc.asp?sponsor=core&z=1727_`
   `00000_2209_dp_04"><img src="image`

> An entity reference was found in the document, but there is no reference by that name defined. Often this is caused by misspelling the reference name, unencoded ampersands, or by leaving off the trailing semicolon (;).

> **The most common cause of this error is unencoded ampersands in URLs** as described by the WDG in "Ampersands in URLs". Entity references start with an ampersand (&) and end with a semicolon (;). If you want to use a literal ampersand in your document you must encode it as "&" (*even inside URLs!*).

> Be careful to end entity references with a semicolon or your entity reference may get interpreted in connection with the following text. Also keep in mind that named entity references are case-sensitive; &Aelig; and æ are different characters. Note that in most documents, errors related to entity references will trigger up to 5 separate messages from the Validator. Usually these will all disappear when the original problem is fixed.

The second tool is a web spider called wget. Wget downloads the number of pages of the URL specified by the user. Wget is a GNU product that can be downloaded from the GNU website (see references). Wget is a network utility to retrieve files from the Web using http and ftp, the two most widely used Internet protocols. It works non-interactively, so it will work in the background, after having logged off. The program supports recursive retrieval of web-authoring pages as well as ftp sites—you can use wget to make mirrors of archives and home pages or to travel the Web like a WWW robot (Wget, 2005). Wget is a well supported program with updated versions available. Also there is an online manual that is required to properly install and run the program.

3.1 Implementation Cont'd

Wget is a free command –line tool that must be downloaded to the computer in order to be used. The version that I used is wget 1.5.3. In the command prompt you must use MS-DOS commands to run the program.

wget –r –l2 –F http://www.gnu.ai.mit.edu/ -o log.txt

This type of command will download the pages of the URL up to level 2 and save the results to a log file in HTML format. The third tool is the WebCred program.
After the user enters the number of validator errors and uses wget to create an HTML log file WebCred analyzes the pages and calculates the credibility score. The scores are from 1-10 points, with 10 being the highest score possible and 1 being the lowest score possible.

3.2 Added Features

In the previous implementation of the WebCred program, the three tools involved in the process had to be used separately from the WebCred form. The changes I integrated to the form, add convenice in the usage of the program. Using Visual C++.NET, I coded a web link to the W3C website on the WebCred dialog. The change gives user access to the validator markup program without having to close the WebCred program. I also added another button object to the form. When the button is clicked a command prompt is opened where the user can follow the steps to run the wget program. Wget is still a command line program and will have to be run that way. This feature was not difficult to implement, it only required another method and the following code syntax: WinExec ("cmd.exe",SW_SHOW) (Chapman, 2002, p. 91). Finally, I added features to enhance the appearance of the program as it is run. I downloaded and installed a file that when the program is run can be seen in the background of the dialog box. This file is a bitmap image of the 10 of diamonds playing card (see references).

The issue of security in WebCred has been addressed in two tracks. Track 1 includes an added feature to the Microsoft Internet Explorer browser. This tool is called DomainSpoofFilter. I downloaded this tool to assist Internet Explorer while searching the Web (see references). This tool is designed to strip hostile redirection from URLs and send users to the legitimate URL, instead of a rogue Web server (Internet Security Systems, 2005). Another aspect of track 1 is downloading another firewall program just in case malicious data escapes the McAfee main firewalls. The program that was chosen in this case was SpywareBlaster version 3.4. SpywareBlaster is freeware and is a useful aid in preventing the installation of spy ware and other potentially unwanted software (javacool software, 2005).

3.2 Added Features Cont'd

The second security track is about the prevention of downloading malicious HTML scripts while spidering the Web. A malicious web developer may attach a script to something you send a web site, such as a URL, an element in a form, or database inquiry. When the web site responds to you, the malicious script comes along, so that it is now on your browser (Cert, 2004). Among the ways you can potentially expose your web browser to malicious scripts are these:

- Following untrustworthy links in web pages, email messages, or newsgroup postings.
- Using interactive forms on an untrustworthy site
- Viewing dynamically generated pages that contain content developed by anyone but yourself (Cert, 2004).

By using a tool to prevent malicious HTML scripts, WebCred provides some protection for its users. I plan to download or implement a tool using wget 's documentation.

4. Evaluation of Program Performance

The enhancements added to the WebCred dialog form make the program easier to use. The link on the form can be clicked, and then the URL validated, and finally the number of errors is returned to be quickly entered into WebCred. The RunWget button can provide quick access to wget after the user has studied the GNU manual for the program. The security tools that have already been downloaded provide confidence that using WebCred is safe. Once the final tool has been implemented to scan the HTML log files, yet another layer of protection will be available to the user. The program could perform even better when the code is optimized for speed. The author of the program did not design the code for speed; therefore the slow processing time is a disadvantage in usage. Once the program is optimized there should not be a problem with using WebCred to aid users in their information searches. Due to time constraints I was unable to implement all of the desired changes to the program. The current program is in a good position to be upgraded further by someone else.

5. Conclusions and Future Work

WebCred is a good tool to use for the following individuals:
- Someone who is somewhat knowledgeable on a topic
- Someone who is curious about the sources of information on the Web.
- Someone who has time constraints in their search for information.

5. Conclusions and Future Work Cont'd

Future work on this project could include: eliminating the domain constraints therefore making WebCred an option for uninformed users.

6. References

[1] W3C Quality Assurance. (2005, July 12). Markup Validation Service v 0.67. Retrieved June 6, 2005, from http://validator.w3.org/

[2] Wget-retrieves files from the Web (n.d.) 2005. Web Authoring. Msc. Wget . Retrieved June 28, 2005 from http://directory.fsf.org/wget.html

[3] Stanford Web Credibility Research (n.d). 2005. Retrieved July 17, 2005, from http://credibility.stanford.edu

[4] Rutherford, LeAne H. (2004, April 22). Caveat Empter: Weighing the Worth of Web Resources. Instructional Development, Spring Quarter, 3, pp 3-5. Retrieved July 19, 2005 from http://www.dumn.edu/~schilton/Courses/Websources.html

[5] Javacool Software (n.d.) 2005. SpywareBlaster 3.4. Retrieved June 30, 2005, from http://www.javacoolsoftware.com/sbdownload.html

[6] Internet Security Systems (n.d.). 2005.Microsoft Internet Explorer Domain URL Spoofing Filter. Retrieved June 1, 2005. from http://www.isss.net/support/product_utilities/domain/spooffilter/

[7] Index of /pub/gnu/wget. (n.d.) 2005. Retrieved June 10, 2005, from http://ftp.gnu.org/pub/gnu/wget/

[8] GNU wget manual. (2005, June 15). 2005 Free Software Foundation. Retrieved June 10, 2005, from http://www.gnu.org/software/wget/manual/wget.pdf

[9] Chapman, Davis (2002). SAMS Teach Yourself Visual C++.NET in 21 days. Indianapolis, Indiana. SAMS Publishing

[10] Cert Coordination Center. (2004, December 7).Frequently Asked Questions About Malicious Web Scripts Redirected by WebSites. Retrieved June 30, 2005, from http://www.cert.org/tech_tips/malicious_code_FAQ.html

6. References Cont'd

[11]   The CARS Checklist (n.d.) 2002. Source
        Evaluation Tutor: CARS. Retrieved July, 20,
        2005 from
        http://www.mnhe.com/science/english/allwrit
        e3/seyler/ssite/se)3/vcs.mhtml

[12]   Brown, Josh, Hickey, Kathy, Pozen, Valerie.
        (2002, July 1) An Educators' Guide to
        Credibility and Web Evaluation. Teaching
        Web Evaluation. Retrieved June 12, 2005,
        from
        http://lrs.ed.uiuc.edu/wp/credibility/index.ht
        ml