Data Mining of Network Logs

Pre-College Internship
Kennedy Space Center (KSC)

Carlimar Collazo
July 28, 2011
Mentor: Henry Yu
About Me

www.theodora.com/maps

UNIVERSIDAD
INTERAMERICANA
DE PUERTO RICO

www.theodora.com/maps
My mission this summer in the Information Security Office (IT-B) was to develop a program which would separate the information in the network logs.
The network logs were getting longer and longer as the advertisements made their way to the most visited web pages. The purpose of this project was to simplify the view of the networks logs (data reduction) so that it would be easier to analyze the information.
Research Techniques

Step 1: Take a look at Uniform Resource Locator (URL) logs.

Step 2: Learn about PERL and its characteristics.
Step 3: Research web page content (including advertisements) and how it works.

Step 4: Get familiar with PERL code.

Step 5: Determine if a link is actual content or an advertisement.

Step 6: Write a software program which separated the content from the ads.
#!perl
#
# 07-jul-2011 carlimar collazo
#Use perl to read a url log delimited with a tab
#
$tab = "\t";
$input_file = @ARGV[0] ;
$output_file = "output.txt";
open (OUT, ">$output_file") || die "Unable to open $output_file: $!
"
; open (IN, "$input_file") || die "Unable to open $input_file: $!
";
#
while (<IN>) {
    chop;
    $curr_line = $;
    ($ip,$date,$action,$w4) = split($tab,$curr_line);
    ($url,$remainder) = split($w4);
    ($d1,$d2,$domain) = split($url);
    #
    print (OUT "$ip$tab$date$tab$action$tab$w4
"
);    #
    print (OUT "ip = $ip
"
);    #
    print (OUT "date = $date
"
);    #
    print (OUT "action = $action
"
);    #
    print (OUT "url = $url
"
);    #
    print (OUT "remainder = $remainder
"
);    #
    print (OUT "domain = $domain
"
);    #
    print (OUT "
"
);    #
    #
    #
    #
    #
}
close(OUT);
close(IN);
Experience with Mentor and Co-workers

My experience with my mentor was awesome. He explained things in a way I could understand. He had confidence in my abilities. This helped me and encouraged me to work hard until I reached the expected goal. My co-workers were always willing to help, explain and respond questions.
Experience with Mentor and Co-workers
Information and Technology Security (IT-B)
Experience with Mentor and Co-workers

We have spirit, yes we do
We have spirit, how 'bout you?!
Knowledge Gained

```cmd
Microsoft Windows XP [Version 5.1.2600]
(C) Copyright 1985-2001 Microsoft Corp.
D:\Documents and Settings\ccollaz1\Desktop>cd Desktop
D:\Documents and Settings\ccollaz1\Desktop>cd "Perl commands"
D:\Documents and Settings\ccollaz1\Desktop\Perl commands>cd ARGV
D:\Documents and Settings\ccollaz1\Desktop\Perl commands\ARGV> example_get_url_fields1.pl urls_201.txt
D:\Documents and Settings\ccollaz1\Desktop\Perl commands\ARGV> example_get_url_fields1.pl url_128.217.135.57_20110707.txt
D:\Documents and Settings\ccollaz1\Desktop\Perl commands\ARGV>
```
The statement of purpose is to analyze network monitoring logs to support the computer incident response team. Specifically, gain a clear understanding of the Uniform Resource Locator (URL) and its structure, and provide a way to breakdown a URL based on protocol, host name domain name, path, and other attributes. Finally, provide a method to perform data reduction by identifying the different types of advertisements shown on a webpage for incident data analysis.

The procedures used for analysis and data reduction will be a computer program which would analyze the URL and identify and advertisement links from the actual content links.

My method of data collection will be based on research about known the advertisement sites and understand the way they are written. In addition, a large part of the data collection will be relying on statistical analysis of the actual log data in order to identify additional advertisement sites.

This project is going to be a big learning experience; the program is going to help the Information Technology Security Office identify the ads that may clutter the actual content data. This would make the process simpler because you would have a program which recognizes the ads right away.

The information contributions are provided by the IT Security Office (IT-B).
Pre College Internship
Sponsoring NASA-Kennedy Space Center (KSC)
Data Mining of Network Logs
Collazo, Carlimar
June 15, 2011

Reviewed by:
Mr. Henry Yu
IT Security Office/IT-B

[Signature]