Seguridad e Integridad de la Información

Práctica 2 Web Crawler

Universidad Iberoamericana

Ivan Enrique Casas Martinez

Abstract: El propósito de la práctica es poder reforzar los conocimientos de bash y para ello se realizó un web crawler el cual puede identificar los correos electrónicos de una página web, lo cual se logró.

Introducción: Hoy en día existen distintas herramientas en la red las cuales realizan distintas actividades, además de esto existen lenguajes como lo es Shell los cuales nos permiten tener un mejor control y entendimiento sobre el funcionamiento total del sistema, con esto podemos asegurar que seremos capaces de realizar procesos con elementos del sistema.

Objetivos:

- Realizar un Web Crawler capaz de identificar los correos electrónicos de una página web
- Aplicar y reforzar conocimiento de bash

Desarrollo:

Para el desarrollo de esta práctica se realizó un script bash el cuál fuera capaz de identificar todos los correos electrónicos que se encontrarán en la URL antares.dci.uia.mx/eortiz/SEGPR20/index.php además de ingresar a los distintos links de la misma para también encontrar los correos que existieran en estos.

Se hizo uso de dsitintos comandos, tales como:

Wget: para poder descargar el contenido de la página

Grep: para poder buscar cierto contenido

Sed: para modificar texto

Cut: para cortar cierto texto

Al final los resultados fueron añadidos a un archivo llamado correos.txt

Código:

```
#Practica 1

#Ivan Enrique Casas Martinez


#!/bin/bash
```

```
wget http://antares.dci.uia.mx/eortiz/SEGPR20/index.php


grep href /home/ic16ecm/Documents/index.php > links.txt
grep @ /home/ic16ecm/Documents/index.php > correos.txt


sed 's/"/;/g' "links.txt" | cut -d ";" -f2 > links2.txt


grep '[A-Za-z0-9]@[A-Za-z0-9].[A-Za-z0-9]' correos.txt | sed 's/"/;/g'
| cut -d ";" -f4 > correos.txt


for linea in $(cat links2.txt)
do
echo "$linea"
wget $linea -o pagina.txt
grep '[A-Za-z0-9]@[A-Za-z0-9].[A-Za-z0-9]' pagina.txt > correos2.txt
for linea2 in $(cat correos2.txt)
do
sed 's/"/;/g' $linea2 | cut -d ";" -f2 >> correos.txt
done
done
```

Conclusiones:

Se logro desarrollar el web crawler reforzando los conocimientos de bash y se pudo observar la utilidad que tienen esta clase de scripts al momento de poder realizar actividades en el sistema.