An Incremental Acquisition Method for Web Forensics

Guangxuan Chen, Zhejiang Police College, China
Guangxiao Chen, Wenzhou Public Security Bureau, China
Lei Zhang, Joint Logistics College, National Defence University, China
Qiang Liu, Zhejiang Police College, China

ABSTRACT

In order to solve the problems of repeated acquisition, data redundancy, and low efficiency in the process of website forensics, this paper proposes an incremental acquisition method oriented to dynamic websites. This method realized the incremental collection on dynamically updated websites through acquiring and parsing web pages, URL deduplication, web page denoising, web page content extraction, and hashing. Experiments show that the algorithm has relative high acquisition precision and recall rate and can be combined with other data to perform effective digital forensics on dynamically updated real-time websites.

KEYWORDS

Digital Forensics, Recall Ratio, Web Crawler, Web Denoising, Web Parsing

INTRODUCTION

As a type of electronic data, website information is one of the legal forms of evidence. The website information that meets certain conditions can be used as the basis for determining the facts of the cases. While web forensics refers to the entire process of capturing, fixing, classifying, storing, and verifying Web sites and page resources using scientific methods. In addition, it sometimes includes forensics of the user’s browsing behavior and related data. The existing Web server forensics is mainly aimed at the attack events of Web pages and illegal services provided by the Web sites, such as pornographic websites, gambling websites, illegal articles websites, and so on. As a branch of general digital forensics, Web server forensics requires relatively complex forensic as it generally involves the analysis of server & client machines, different types of OS and network devices (Chen et al., 2013).

Due to the cross-platform characteristic and distributed framework of Web applications, investigators should have a deep understanding about the systems and platform on which the Web application runs. The analysis is usually started from the logs on the server, and the contents of the database are also the key objects. The difference between the online database and the offline database needs more attention during the forensics process. The trace information of attacks to the Web server usually can be found on the main server and client, even the application server and database server. Therefore, it is necessary to perform forensic analysis on the web server, client, application server, and database server at the same time, and compare the evidence information with each other to ensure that the information is true and reliable (Chen et al., 2015).

DOI: 10.4018/IJDCF.2021110116

Copyright © 2021, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.
For web content collection, web crawling is usually adopted. Web crawlers are usually used to obtain various information and data on the network according to certain rules or purposes. At present, the most popular web crawlers are general purpose web crawler and focused web crawler. The former is also named scalable web crawler. Its crawling object extends from certain seed URLs to the entire Web and it mainly collect data for portal site search engines and large Web service providers. The latter is also named topical crawler, which selectively crawl pages related to the pre-defined topics. Compared with the general web crawler, focused web crawler only needs to crawl the pages related to the theme, which greatly saves hardware and network resources. The saved pages are also updated quickly due to the small number, and can also meet the needs of some specific fields (Chen et al., 2020).

Due to the time-sensitive and fast-changing nature of web page data, the traditional forensics method can no longer meet the needs of law enforcement agencies. While incremental acquisition can provide a solution. The incremental acquisition system can run multiple crawler tasks at the same time. These incremental crawlers refer to crawlers that take incremental updates to downloaded web pages or crawl newly generated or changed web pages. Incremental crawlers periodically revisit the website according to different incremental strategies. Common strategies include: statistical update method, individual update method, classification-based update method, etc. Data acquisition oriented to a specific website can be defined as a crawler task. The website can be a portal website, a forum, or a data interface of some social networking sites. The data acquisition can be file download or request access to the server interface. In the incremental acquisition system, each crawler task is independent of each other and has its own separate seed list, crawling thread, link queue, filtering rules, etc. Every crawler task has a life cycle, including task creation, operation, completion and uninstallation. In the life cycle of a task, the acquisition system is responsible for task creation, monitoring, management and scheduling. The system will create a separate management directory for each task to store all the data required by a crawler task. When the task starts to run, the system reads the seed file from the task directory, reads the configuration information, and creates threads and queues for the task. During operation, the system is responsible for allocating links to threads and queues, parsing pages, and storing data. When the task ends, the system clears the thread and queue information from the memory and updates the management directory. Multiple crawler tasks are processed in parallel in the acquisition system without interfering with each other.

Aiming at the new problems encountered in website forensics, this paper proposes an incremental website data collection method and forensic model, and implements a credible webpage forensics scheme to ensure the credibility of the collected content. In the process of forensics, cooperating with the rapid filtering of massive web data can improve the efficiency of forensics, realize the rapid extraction and analysis of data released by website users, and help to respond to the criminal investigation and evidence collection.

RELATED WORK

Website Forensics

Website forensics is one of the branches of digital forensics. It is a forensic technology developed corresponding to web security threats and mainly targets web-related criminal activities, such as phishing, XSS attacks, pornographic websites, gambling websites, and so on. Zhong et al. proposed a web forensics technology based on web data streams (Chen et al., 2013). It analyses the network data packets and conduct forensics using sniffing technology. The traditional methods can be used to conduct the forensics of the server and the client after fixing the data storage device. The main difficulty is that each relevant application needs to be analyzed in detail, such as log information extraction. Xiong et al. proposed a dynamic real-time forensics method for web pages, and encrypted the obtained data package to ensure the validity of the potential evidence and through the third-party security center to preserve the reliability of the evidence package (Xiong et al., 2016). F. Facca et
al. proposed to analyze the corresponding user and session ID and use it as instruction information to collect information records about the user and its session in the web server and application server logs. Wu et al. proposed the web server forensics technology which conducts forensics mainly from the website server, the suspect’s computer and network stream (Wu et al., 2015). Among them, the forensic work on the server side can start from the querying log files, configuration files, etc. from the web server, application server and database.

**Website Information Acquisition**

In the process of website forensics, it is usually necessary to collect website data by all means, such as web page information. In an ideal situation, the collected information should be consistent with the information on the web. However, the dynamic, heterogeneous and complex nature of the web determines that the collected information may become outdated in a relatively short period of time. It is unrealistic, and all the authors can do is tried to approach this ideal as much as possible. Common methods include topic-based collection methods and incremental collection. The main idea of topic information collection is to judge the topic relevance of the collected pages, i.e., predict the relevance of the URL to be collected with the topic and accordingly decide the direction of the collection according to the topic characteristics of the collected content during the acquisition process. The focused crawling solution proposed by Chark et al., determines the nature of the collected pages through the Hub/Authority feature, and uses the Linkage Locality and Sibling Locality features of the topic information to determine further collection border based on the existing topic page (categorized) (Cao, 2018). Here, the Linkage Locality feature indicates that if the content of a page is related to a topic, it has a greater possibility of linking to other pages with the same topic. The Sibling Locality feature points out that among the hyperlinks contained in a page, if the number of evaluated hyperlinks related to the theme is greater, the remaining hyperlinks may be more likely relevant to the topic. The predictive factors considered by the Intelligent Crawling scheme proposed by Aggarwal et al. include: content (analysis of the relevance of the collected page text); URL Token (check whether the URL address contains keywords of the topic); Link (the reference relationship of the Hyperlink between the collected URLs and URLs to be collected); Sibling (relevance statistics of URLs on the same page) (Cooper et al., 2004). The idea of this scheme is that there is no need to classify topic pages in advance, but start searching directly with a small number of keywords, which has better applicability. However, the effect of this scheme is sensitive to the subject keywords and is often less effective than method based on constructing classifier, and the effect can only be reflected when the collection page reaches a considerable scale. Ma Liang et al proposed an intelligent web information collection scheme oriented to chinese topic. This scheme considers the organizational characteristics of the topic information when evaluating the topic relevance of the collected pages, and integrates the page title, paragraph title (usually determined by special fonts), and Anchor text (the explanatory text of the referenced URL), and several other factors that have a higher value for evaluation. These factors are given a relatively high weight coefficient so as to improve the accuracy of the evaluation results (Dai, 2015).

Compared with periodic information collection, incremental information collection can greatly reduce the amount of data collection and thus greatly reduce the time and space overhead, so it has become the first choice for the actual collection system. However, while the incremental information collection reduces the time and space overhead, it also increases the complexity and difficulty of the algorithm. At the same time, it also faces new problems, such as how to allocate the system’s collection capacity according to the changes of the pages. At present, there are several common incremental web crawlers oriented to special objects, such as forum oriented and e-commerce website oriented web crawler. In addition, a denoising based web crawler proposed a solution for the problem that the classic algorithm is too sensitive to the value generated by the text. However, the current incremental crawlers often target a certain field and the network incremental crawling is relatively rough, which cannot meet the actual refined data collection requirements (Heydon et al., 1999). On the basis of
analyzing the OutbackCDX and UKWA-Heritrix systems, Gao and Bai proposed an incremental acquisition scheme based on OutbackCDX and UKWA-Heritrix. This scheme can realize the fixed title and incremental acquisition without restarting the crawler program, and uses the ISO standard format WARC (Web ARCHive) as the storage format to realize the direct data exchange between different types of systems (Gao & Bai, 2020). Yang realized the incremental acquisition of Internet website archives by adding additional tools to the acquisition system, achieving the goals of reducing server operating load, network bandwidth occupation, website archive storage space and improving the display quality of collected resources (Yang, 2020).

THE THEORETICAL BASIS OF INCREMENTAL ACQUISITION TECHNOLOGY

Incremental acquisition technology, in fact, is to build a model for each web page, and calculate their changes. Due to the huge number of the webpages and their obvious aggregation natures, the modeling usually aims to each website and estimates the time variation of it in actual process. Then, set up an incremental information collector for the websites with the same type of changes.

Generally speaking, by judging the collected web pages through experimental means, the authors can find the changing rules of the web pages. This method can be extended to the entire website and the entire Internet. The changing process of web pages can be regarded as a Poisson process (Haight, 1967; Meng et al., 2006; Huang, 2006). The Poisson process is a basic independent incremental process that accumulates the number of occurrences of random events (Meng et al., 2006; Meng, 2015). Starting from a certain time 0, $X$ is used to indicate the number of changes of a certain web page at time $t$. Each change of the web page is independent and identically distributed, and the change frequency is $\lambda$, according to the definition of the Poisson process, it has

$$\text{Pr}\{X(s+t) - X(s) = k\} = \frac{(\lambda t)^k}{k!} e^{-\lambda t} \quad (1)$$

For equation (1), when $k = 1$, that is the time $T$, the next time when the webpage changes, the density function of the interval between its two changes is shown in equation (2) (Meng et al., 2006; Li, 2011).

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t}, & t > 0 \\ 0, & t \leq 0 \end{cases} \quad (2)$$

Here, the authors use two definitions to measure the status updates, freshness and age of web page (Meng et al., 2006; Huang, 2006). If at a certain time $t$, the actual content of the web page on the web server is the same as the content saved locally, then the web page is regarded as up-to-time, and give the value 1. Otherwise, give the value 0. As equation (3) shows

$$F(e; t) = \begin{cases} 1, & \text{if } e \text{ is up-to-time} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

For the age of a webpage, as shown in the equation (4), if the webpage is up-to-time, then its age is given the value 0. Otherwise, its age refers the period from the time of the last update to the
current time t. The longer the update cycle time, the older the page may be. Here, $LMT(e_i)$ is the time of the last update of webpage $e_i$ before time $t$.

$$A(e_i, t) = \begin{cases} 0, & \text{if } e_i \text{ is up-to-time} \\ t - LMT(e_i), & \text{otherwise} \end{cases}$$  

(4)

For a set $S$, its average fresness and average age are shown in (5) and (6) (Meng et al., 2006; Marrington, 2013).

$$F(S, t) = \frac{1}{N} \sum_{i=1}^{N} F(e_i, t)$$

$$A(S, t) = \frac{1}{N} \sum_{i=1}^{N} A(e_i, t)$$

(5)

The incremental collection system maintains a webpage collection $S$, which is concerned with the average freshness and average age of $S$ over a period of time. At this time, the average value of in time domain can be measured as

$$\overline{F(S)} = \lim_{t \to \infty} \frac{1}{t} \int_{0}^{t} F(S, t) dt$$

$$\overline{A(S)} = \lim_{t \to \infty} \frac{1}{t} \int_{0}^{t} A(S, t) dt$$

(6)

As for a single webpage, its average freshness and average age over a period of time can be obtained by through equation (1) to equation (4). Assuming that the average change frequency of a webpage is $\lambda$, then the probability that it will change in the interval $I = (0, t)$ is

$$\Pr(T \leq t) = \int_{0}^{t} f_T(t) dt = \int_{0}^{t} \lambda e^{-\lambda t} dt = 1 - e^{-\lambda t}$$

(7)

Therefore, the average freshness and average age of webpages in interval $I$ are

$$E[F(e_i, t)] = 0 \cdot \Pr\{T \leq t\} + 1 \cdot \Pr\{T \leq t\} = e^{-\lambda t}$$

$$E[A(e_i, t)] = t \cdot \left(1 - \frac{1 - e^{-\lambda t}}{\lambda t}\right)$$

(8)

The above definition establishes a basic mathematical model for the changing process of webpages, and defines two important indicators for evaluating the freshness and age of the webpages. The average freshness and average age in interval $I$ can be derived according to equation (5), (6) and (8).
DESIGN OF INCREMENTAL ACQUISITION

Due to the strong timeliness and rapidity of changes of the web page, web page forensics needs to go further on the basis of data capture and static analysis. According to the Poisson process of the changes of the web page, the authors need to set up a model for each web page to calculate their change rule. Generally speaking, for a website, the number of web pages is often large and the web pages have an obvious agglomeration. Therefore, the model here is oriented to the whole website and is used to estimate the time change rule of the website. Then, an incremental data collector is designed for the same type of changing website, which can be suitable for dynamic, real-time, orderly, efficient and reliable web page forensics.

On this basis, the authors designed a web crawling model. Here, the scrapy crawler framework is adopted to crawl web content. And several key parameters such as web page refresh time and web page fingerprints in the website are fully considered in the incremental acquisition algorithm. The incremental crawling algorithm of the page is designed as follows:

Step 1: Put the initial URL seed and meta information into the model. The web page refresh time and web page fingerprint in the URL meta information can be left blank. If there is no refresh time and webpage fingerprint in a URL meta information, then it can be determined as URL seed address and set the initial value automatically.

Step 2: Collect webpage based on the address list. For each URL in the list, use the HEAD method of the HTTP protocol to request the HEAD information of the webpage corresponding to this URL, obtain the last modification time and the length of the webpage, and compare them with the last modification time and the length of the webpage in the URL meta information. If there is a difference between the two, then start collecting, otherwise it means that the webpage has not changed and no collection is needed.

Step 3: Resolve the collected webpage information and generate URL meta information for each URL.

Step 4: Calculate the web page collection time based on the current meta information and the previous meta information of the URL, update current meta information of the URL, and store it in the URL priority queue according to the web page collection time.

Step 5: Get the URL list from the URL priority queue and then start to collect the data.

By actively accessing the target web page and initiating a request to the data resource server corresponding to the requested domain name, the authors can quickly obtain the surface data, interaction data and underlying data of the requested page.

EXPERIMENTS AND ANALYSIS

Experiment Design

In this work, the authors design experiments to test the acquisition performance of the incremental algorithm, and use 9 server nodes (the hardware configuration and deployment environment are shown in Table 1) with the same configuration to perform incremental acquisition tests on the selected 9 target sites. These websites are different in the nature of the main body, the programming language used, and the functional purpose. And they are now completely taken over from the former website server and redeployed on the new servers. It has been slightly updated according to the way it used to work: the website updates the page content every 10 minutes, automatically adds a news message, or changes a headline, or inserts a picture, or adds a complete page. The initial acquisition period is set to 5 minutes. The parameters of the tests are shown in Table 2.

Here, the authors selected two popular open source incremental crawlers, Octopus collector (Cui & Liao, 2016; Zhang, 2019) and SpiderInfo (Han et al., 2019; Wang et al., 2019), to compare the algorithm proposed in this article. Octopus collector is an open source tool developed by the Java language. It not only provides the function of web collection, but also can analyze the pages
and build index. It provides all the tools needed to run a search engine. The SpiderInfo project is an open source and scalable crawler project based on the Java language, which has better scalability than Octopus collector. Using the scalability of SpiderInfo, developers can extend its various components, customize their own crawler tasks, and implement their own crawling logic.

**Results and Discussion**

In the experiments, the precision and recall rate are used to evaluate the performance of the incremental acquisition algorithm. Precision and recall rate are two measures widely used in the field of information retrieval and statistical classification to evaluate the quality of results. The precision is the ratio of the number of related documents retrieved to the number of total documents retrieved, which measures the accuracy of the retrieval system; the recall rate refers to the ratio of the number of relevant documents retrieved to the number of all relevant documents in the document library.

Generally speaking, precision refers to how many of the retrieved items (such as documents, web pages, etc.) are accurate, and recall refers to how many accurate items are retrieved. Precision and recall are defined as follow:
Here, TP, FP and FN refer to the number of new pages in the collected sets, the number of old pages collected, and the number of new pages uncollected in the site respectively. The values of the precision and recall are between 0 and 1. The closer the value is to 1, the higher the precision or recall rate.

Table 3 shows the acquisition results of the experiments. It can be seen from the table that the average precision and average recall of the incremental acquisition on the 9 websites are 96.95% and 96.76%, which are relatively close to 1, indicating that the algorithm has a high precision and recall rate.

Figure 2 and Figure 3 show the precisions and recalls of the incremental acquisition on the 9 target websites. As can be seen from the comparison, the scheme presented in this paper has higher accuracy than Octopus Collector and SpiderInfo on site01, site02, site03, site04, site05, site06, site07, site08, and higher recall ratio on site01, site02, site03, site04, site05, site06, site08. These
Table 3. The results of incremental acquisition of the proposed method

<table>
<thead>
<tr>
<th></th>
<th>Number of web pages collected</th>
<th>Number of new pages on the site</th>
<th>Number of new pages in the collected sets</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>site01</td>
<td>681</td>
<td>672</td>
<td>654</td>
<td>0.9604</td>
<td>0.9732</td>
</tr>
<tr>
<td>site02</td>
<td>810</td>
<td>800</td>
<td>788</td>
<td>0.9728</td>
<td>0.9850</td>
</tr>
<tr>
<td>site03</td>
<td>947</td>
<td>963</td>
<td>940</td>
<td>0.9926</td>
<td>0.9761</td>
</tr>
<tr>
<td>site04</td>
<td>1088</td>
<td>1080</td>
<td>1042</td>
<td>0.9577</td>
<td>0.9648</td>
</tr>
<tr>
<td>site05</td>
<td>1305</td>
<td>1312</td>
<td>1287</td>
<td>0.9862</td>
<td>0.9809</td>
</tr>
<tr>
<td>site06</td>
<td>1421</td>
<td>1450</td>
<td>1394</td>
<td>0.9810</td>
<td>0.9614</td>
</tr>
<tr>
<td>site07</td>
<td>1480</td>
<td>1500</td>
<td>1441</td>
<td>0.9736</td>
<td>0.9607</td>
</tr>
<tr>
<td>site08</td>
<td>1547</td>
<td>1600</td>
<td>1531</td>
<td>0.9897</td>
<td>0.9569</td>
</tr>
<tr>
<td>site09</td>
<td>1965</td>
<td>2005</td>
<td>1904</td>
<td>0.9690</td>
<td>0.9496</td>
</tr>
</tbody>
</table>

Figure 2. Precision of the incremental acquisition experiments
experiments show that compared with the other two models, the model proposed in this article has better incremental acquisition performance for different types of websites. In addition, as the number of acquisition increases, the acquisition cycle will fluctuate along the real changing cycle of the webpages. It effectively reduced the amount of redundant data and realized the pressure of the network flow.

Combined with the basic information of the target website obtained through other means, such as routing, path, IP, ICP, server resource distribution and other information, the authors can carry out subsequent data analysis and evidence extraction, and solidify the obtained evidence, and at the same time generate digital fingerprints and the time stamp of the National Time Service Center and send them to a third-party security center with judicial effect for preservation to ensure the legality and uniqueness of the evidence.

CONCLUSION

As for the large number of illegal and crime-related websites, the traditional static web page collecting and data analysis are no longer applicable. In the face of frequently updated web page, a real-time and effective means of forensics is needed. This paper proposes a dynamic, efficient and reliable website incremental acquisition model. By actively accessing the target web site and initiating a request to the

Figure 3. Recall ratio of the incremental acquisition experiments
data resource server corresponding to the requested domain name, it can quickly obtain the surface data, interaction data and bottom layer of the requested web page data. In addition, the incremental acquisition algorithm based on Poisson distribution avoids the waste of time and data redundancy caused by repeated collection of unchanged web pages during the forensics process. Through the design experiment, the target website is collected incrementally, and it is found that the algorithm has a better accuracy rate and recall rate. Combined with other data obtained from the website, an effective webpage forensics can be performed.

The incremental acquisition program proposed in this article is not a general-purpose crawler. It is aimed at refined data acquisition requirements. However the current acquisition is not efficient to complex website. It is necessary to design an optimized crawler algorithm if the authors want the crawler to be more efficient. Therefore, the further research will focus on how to improve the crawling efficiency of crawlers. Furthermore, in order to deal with the anti-crawler mechanism of the website, research on crawler agents will be launched.

In addition, hacker’s attack is also a problem that must be taken into account. The massive meaningless websites generated by the owner will interfere with the forensics. Therefore, these factors will be considered in subsequent studies.

**ACKNOWLEDGMENT**

In this paper, the research was sponsored by the Basic Public Welfare Research Program of Natural Science Foundation of Zhejiang Province under Grant No. LGF19F020006 and NSFC program under Grant No.62002362.
REFERENCES


Guangxuan Chen is an associate professor of Key Laboratory of Big Data, Ministry of Public Security. He received the Ph.D. degree in digital forensics in 2014. From 2016 to 2019, he was a postdoctoral fellow at the Institute of Software, Chinese Academy of Science. And he was the visiting scholar with the University of Lausanne, Switzerland from 2016 to 2017. His research interest includes the cloud computing, cyber crime investigation and digital forensics.

Guangxuan Chen graduated from Universidad Carlos III de Madrid, Spain with a master’s degree. His research interests include cybercrime investigation and digital forensics.

Zhang Lei, associate professor of the Joint Logistics College of National Defence University, graduated from China University of Geosciences (Beijing) in 2009 with a master’s degree in computer application technology. Her research interests include digital forensics, cloud computing.

Qiang Liu is an associate professor of Zhejiang Police College. His research interests include digital forensics, bigdata, and so on.