ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

Crawling Web Information for Forensic

Sameer Jadhav

M.E. Student, Dept. of Computer Engineering, MMCOE, Savitribai Phule Pune University, Pune, India.

ABSTRACT: The World Wide Web is full of information. If you want to know something, you can probably find the information online through search engines. Web Crawler is an Integral part of the search engine that go around the URLs to gather the information from the Web. Focused Crawlers are used to retrieve topic specific information from the web. They try to find the information by overlooking the irrelevant links which are not based on specified topics. The Deep Web is a network of interconnected systems, not indexed, having a size of around 500 times higher than the current web and is continued to increase steadily. However, due to the exponential growth of the Internet, large volume of web resources and the ever changing nature of deep web, achieving wide coverage and high efficiency is a challenging issue. In this paper, we have applied the concept of focused crawling for information retrieval of Organizations which are involved in illegal activities. The aim is to develop a System which maintains the database of such Organizations involved in illegal activities. System also consists of a pre-query component in order to filter irrelevant input and a post-query component to filter out irrelevant retrieved information. Gathering all such information through manual search is an impossible task and hence this paper gives a design of a focused crawler that can gather all such information. This continuously updating database will provision forensic Organizations to investigate and analyze the evidence.

KEYWORDS: Computer Forensic, Web Crawler, Deep Web**,** Focused Crawling, Pre-query, Post-query.

I. INTRODUCTION

The surface Web contains an estimated 2.51 billion documents, growing at a rate of 7.51 million documents per day. The Deep Web (or Invisible web) number of non-indexed web sites is the set of information resources on the World Wide Web not reported by normal search engines. Deep Web is estimated to be 450 to 550 times larger than the surface web of indexed, searchable web sites.Today, most of the popular search engines have provided us with facilities to retrieve any intelligence on the Internet. When user tries to search for any information, they usually focus on some specific topic or person. Search engines use Web crawlers to collect information available online. Web crawlers are the tools that keep on following the hyperlinks to gather information. Rather than collecting all the available data on the Web, focused crawler selectively download webpages that are relevant and satisfy some specific property.

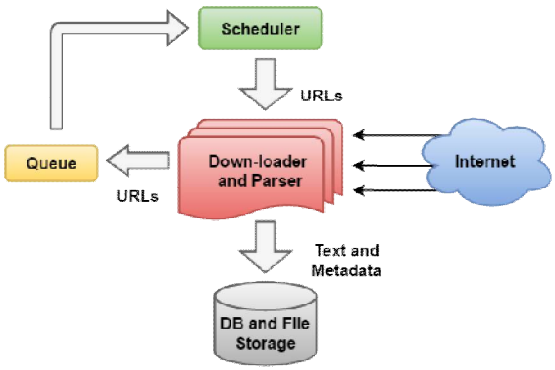

Fig.1. System Architecture of a Web crawler

A crawler is a program that visits Web sites and reads their pages and other content in order to create entries for a search engine index. The leading search engines on the Web all have such a program, which is also known as a

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

"spider" or a "bot." Entire sites or specific pages can be selectively visited and indexed.The concept of focused crawling was introduced in [2]: a focused crawler can seek, acquire, and index webpages on a specific set of topics that represent a limited portion of Web. Focused crawling approach minimizing resources spent fetching pages on other topics, and helps to keep the data collected by the crawler more up-to-date.

There are some online national Web accumulations available that preserve webpages that are related to their nations [3],[4]. Most of such accumulations compromise information ofhistory, profundity and civilization of the countries. Thereis an instant need for a database of computer forensic toknow which Organizations where previously or currentlyparticipating in the illegal activities.

This will not only serve the computer forensic analyst but will be beneficial for other Organizations to do businessdeals. We have applied this concept of focused crawling for information retrieval of Illegal Activities involved withspecified Organization.

Websites of news around the sphere will be required sectionthat needs to be crawled to get the course of study of suchOrganizations. This search interface of Organizations willgive Computer Forensic Analyst a service who desire toinvestigate a case.

II. RELATED WORK

Searching and indexing of the Web by search engines or other applications can be achieved by using Web crawlers.Web Crawlers are the programs that visits Web sites andreads their pages and other content in order to create entries for a search engine index. The major search engines on the Web all have such aprogram, which is also known as a spider or a bot. Crawlerswork by recording every hypertext link in every page theyindex crawling. It is challenging to locate the deep webdatabases, because they are not registered with any searchengines and keep constantly changing.

To direct this problem, previous work has proposed two types of crawlers, Generic Crawlers and Focused Crawlers:Generic Crawlers fetch all searchable forms and cannot focus on a specific topic.Focused Crawler selectively seeks out webpages that are relevant to a pre-defined set of topics [1]. Rather than collecting and indexing all available Web documents, a focused crawler analyses its crawl boundary depending upon the conditions specified by the user. It finds the links that are likely to be most relevant for the crawl, and avoids irrelevant regions of the Web. An up-to-date review of various crawling algorithms for focused crawler is given in [5] and [6]. We need to specify a set of seed URLS to the focused crawler. The seed URLs are the starting URLs from which a crawler starts the process of crawling.

Besides efficiency, quality and coverage on relevant deep web sources are also difficult. Crawler must produce a largequantity of high-quality results from the most relevant contentsources. Thus it is crucial to develop smart crawling schemethat are able to quickly discover relevant content sourcesfrom the Deep Web as much as possible.Traditionally language specific crawlers [7], [8] are used tocollect webpages written in their national language. To ourknowledge [9] was the first who adopted focused crawler tocollect webpages written in a specific language and called itlanguage specific Web crawler.There are a number of open source crawlers available in different programming languages. They can be modifiedaccording to user requirements.

A. Some of the open source crawlers are briefly discussed below.

- HTTrack [10] is a tool for the users unfamiliar withWeb crawling and it provides them with a GUI. It isdesigned to create mirrors from existing websites. It canfollow links that are generated with basic JavaScript andinside Applets or Flash. The drawback with HTTrack isthat it does not have integration with indexing system,so the data obtained cannot be used for further processing.
- Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang,Hai Jin [1] provides a two-stage system, for the SmartCrawler, for efficient aggregating deep web interfaces.Initially, Smart Crawler carry out site-based searchingfor focused pages with the support of search engines,holding up from going to an extensive quantity of pages.
- In [11] major focus was on the fact that effective flowscan be used to produce highly effective results on web.The filters incorporated with the used algorithms in thepaper are well effective and high performance for websearch, trim the network traffic and crawling costs.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

- The work in [12] focused on using query preprocessing using fuzzy logic and also suggested that the query-basedexecution is based on the query range, a measure ofthe query particularity. The query scope is definedusing probabilistic propagation mechanism on top of thehierarchical structure of concepts provided by WordNet.
- Also [13] focused that forecasts can be generated beforethe retrieval procedure takes place, which is morepractical than current approaches to query performanceprojection. The approach was calculated with the linearand non-parametric correlations of the forecasts withAverage Precision.
- The work in [14] used a model that focused on selectivecutting framework for ensuring strong retrieval, byappropriately setting the cutting parameters of Wand ona per-query basis, before re-ranking the results using alearned model.

This work uses the concept of focused Web crawler. The similar work done in the literature is studied in detail bythe authors. In addition, the open source Web crawlers were explored. We aimed at applying these concepts on the topicof: Organizations involvement in illegal activities information retrieval.

III. PROPOSED ALGORITHM

A. *DESIGN CONSIDERATIONS:*

- A subset of the web consists of Organizations Information of Illegal Activities.
- User queries a valid Organizations name/Person name as input for retrieving specific information.
- Complex problem is divided into small subsets increasing the system performance. Amonolithic classifier is replaced with Hierarchy of classifiers.

B. *DESCRIPTION OF THE PROPOSED ALGORITHM:*

Crawling has many constraints that needs to be consideredwhen developing an application, the biggest being a constrainton the resources. Each crawler has a limited set of resourcesand bandwidth allocated to it. Effective utilization of theseresources is must to crawl maximum number of webpages.

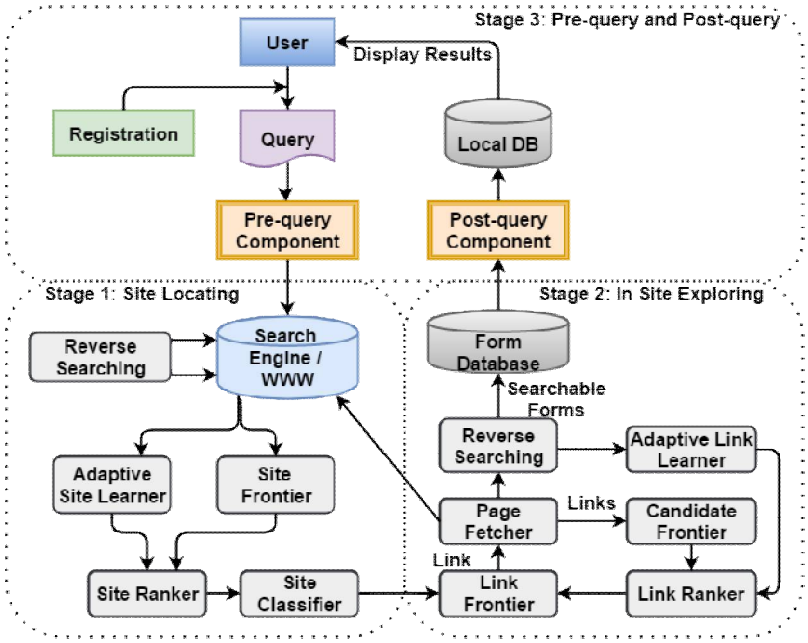

Fig.2. System Architecture of CWIF

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

While crawling we need to ensure that the crawled webpageshave not been modified or have not expired. There is a limit on the number of webpages that can be visited by a Webcrawler. The order in which the webpages are to be traversed also needs to be defined.

In proposed system, User does the registration in the application and then provides an input text to the system.Then the pre-query component system checks the relevance of the input. In the Site Locating Stage, the subset of the entire Webis chosen that is to be visited. Web pages are traversed andt he system does Reverse Searching based on existing search algorithms to find seed sites. We also specify the limit byt hreshold value to search results. As these webpages are traversed all the URLs encountered on the webpages visited re collected and stored away for later traversal.

Then the links of the seed sites are processed by the "In Site Exploring" stage. It represents the constraint based data extraction. Over that links Nave Bayes classifier is enforced for to categories the links or sites. After, on the basis of ranking, ranking of links or sites are accomplished. Then the result is processed by the post query componentto filter out the irrelevant information. And the final output is stored in the database and displayed to the user.

IV. PSEUDO CODE

Step 1: User logins to the Proposed System.
Step 2: User inputs a search query.
Step 3: Pre-Query processing.
Step 4: Fetch deep sites by reverse searching.
Step 5: Incremental site prioritizing.
Step 6: Classify links using Nave Bayes.
Step 7: Extract domain specific forms.
Step 8: Post-Query processing.
Step 9: Output Highly Relevant Information.
Step 10: End.

V. SIMULATION RESULTS

To evaluate the performance of our crawling framework, we compare CWIF to the SCDI (site-based crawler for deep web interfaces) and ACHE (An adaptive crawler for locating hidden-web entry points). The dataset used to analyze the system is TEL-8 dataset of the UIUC repository. The TEL-8 dataset contains 447 deep web sources with 477 query interfaces, because a source may contain multiple interfaces. We have implemented CWIF in Java and evaluated our approach over 12 different domains described in Table 1.

TABLE 1: Twelve domains for experiments.

Domain	Description
Airfare	airfare search
Apartment	property search
Auto	used vehicles search
Book	books search
Hotel	hotel search
Job	job search
Movie	movie titles and DVDs search
Music	music CDs search
Rental	car rental search
Route	map and airline search
People	celebrities search
Product	household appliances search

Figure 3 shows that CWIF finds more relevant deep websites than ACHE and SCDI for all domains

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

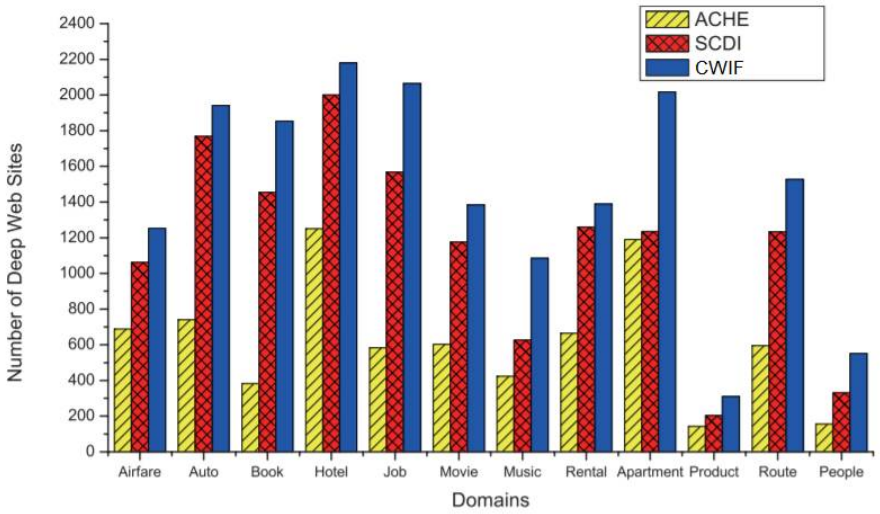

Fig.3. The numbers of relevant deep websites harvested by ACHE, SCDI and CWIF.

The aim of the study is to develop a focused Web crawler that aspires selective information retrieval from the Internet.The proposed crawler is meant to retrieve information aboutOrganizations involved in illegal activities. The problem canbe precisely defined as to develop a focused crawler that canbe used to extract information of Organizations which areinvolved in illegal activities using news websites of variouscountries as seed URLs.

A. ***THE FOUR MAIN OBJECTIVES THAT COME OUT ARE:***

- To crawl, extract and index the data for Organizations illegal activities up to a predefined depth of crawling.
- To prepare a set of keywords specific for guiding the crawling process.
- To continuously update the collected information using a proper revisit policy.
- To add a pre-query and post-query component in the system.

The database that will be developed will have all theinformation of Organizations involvement in illegal activities.

B. ***THE DESIRED DELIVERABLES OF THE WORK WILL BE:***

- A database of the Organizations involved in illegalactivities.
- A focused crawler that will gather and index the desired information. In addition, it will keep the indexedinformation up to date.
- User Interface for extracting information of interest by the users.

VI. CONCLUSION AND FUTURE WORK

This paper discusses the need for building an online database of Organizations involvement in illegal activities. This research presents a novel idea of creating an application for retrieving information from deep web by using smart crawler technique which consists of 3 stages. Manual searching of such information is not feasible and hence we need to build a focused crawler that can crawl websites with news domain. Keyword match based approach is used for determining the relevancy of webpage. Depth of the webpage is used as stopping criteria. Efficient crawling is achieved by restricting the crawling process only to URLs that fall in the domain of parent seed URL. The developed system will provision forensic Organizations to investigate and analyze the evidence. And also will be beneficial for other Organizations to do business deals. Various challenges and issues related to general focused crawler andCrawler4j are also highlighted.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

In future various applications can be built depending upondata extracted for better decision making. Thus to concludethe focused crawler developed so far has the capability tocrawl websites and download data automatically. This can befurther scaled to keeping it rigorously running and to optimizeperformance, thereby making the crawling procedure muchfaster and efficient.

REFERENCES

1. Feng Zhao, Jingyu Zhou, Chang Nie, Heqing Huang and Hai Jin, SmartCrawler: A Two-stage Crawler for Efficiently Harvesting Deep-Web, Interfaces. IEEE Transactions on Services Computing Volume: PP Year: 2015.
2. SoumenChakrabarti, Martin van den Berg 2, Byron Domc, Focusedcrawling: a new approach to topic specific Web resource discovery, Published by Elsevier Science B.V. 1999.
3. Warp.ndl.go.jp, 2011. [Online]. Available: http://warp.ndl.go.jp. [Accessed:02- March- 2015].
4. Was.nl.sg, Web Archive - National Library Board, Singapore, 2013.[Online]. Available: http://was.nl.sg. [Accessed: 02- Feb- 2015].
5. G. Pant, P. Srinivasan, and F. Menczer, Crawling the web. In WebDynamics, pp. 153-177, Springer Berlin Heidelberg, 2004.
6. FilippoMenczer, Gautam Pant and PadminiSrinivasan, Evaluating Topic Driven Web Crawlers, 2011.
7. E. Srisukha, S. Jinarat, C. Haruechaiyasak, and A. Rungsawang, Naïve bayes based language-specific web crawling, in Proceedings of the 5thInternational Conference on Electrical Engineering/Electronics, Computer,Telecommunications and Information Technology,Vol.1 pp. 113-116, IEEE, 2008.
8. P. Tadapak, T. Suebchua, and A. Rungsawang, A machine learningbased language specific web site crawler, in Proceedings of the 13th International Conference on Network-Based Information Systems (NBiS), pp. 155-161, IEEE, 2010.
9. T. Tamura, K. Somboonviwat, and M. Kitsuregawa, A method forlanguage-specific web crawling and its evaluation, Systems and Computers in Japan, vol. 38(2), pp. 10-20, 2007.
10. Httrack.com, "HTTrack Website Copier - Free Software Offline Browser(GNU GPL)", 2014. [Online]. Available: http://www.httrack.com/. [Accessed: 08- Feb- 2015].
11. Rahul kumar, Anurag Jain andChetan Agrawal,Survey of Web Crawling Algorithms, Advances in Vision Computing: An International Journal (AVC) Vol.1, No.2/3, September 2014.
12. Ben He andIadh Ounis, Inferring Query Performance Using Pre-retrieval Predictors, Department of Computing Science University of Glasgow fben,ounisg@dcs.gla.ac.uk.
13. Oren Kurland, Anna Shtok, David Carmel, and Shay Hummel,A Unified Framework for Post-Retrieval Query-Performance Prediction,ICTIR 2011
14. Steven M.Beitzel, Eric C. Jensen, Abdur Chowdhury and Ophir Frieder,Varying Approaches to Topical Web Query Classification, SIGIR07, July 2327, 2007, Amsterdam, The Netherlands, ACM.
15. Vassilis Plachouras, Iadh Ounis. Springer-Verlag,Dempster-Shafer theory for a query-biased combination of evidenceon the Web, Berlin Heidelberg 2014.

BIOGRAPHY

Sameer Ravikant Jadhav is currently pursuing Master's in Computer Engineering from Marathwada Mitra Mandal's College of Engineering (MMCOE), Savitribai Phule Pune University. He received Bachelor of Computer (B.E.) degree in 2013 from MMCOE, Pune-52, India. His research interests are Information Retrieval and Data Mining.