IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 18, Issue 1, Ver. II (Jan – Feb. 2016), PP 31-36
www.iosrjournals.org

An Efficient Strategy to Reduce and To Restrict Malicious Attacks on the Web with Web Sense

Dr V.R.Elangovan[1]
[1]*Department of Computer Applications, A.M.Jain college Chennai,Tamilnadu India*

Abstract*: As the growth of Information Technology and Communication has led to vivid change, people can connect with each other globally at anywhere and in anytime through search engines. These search engines has also tremendously increasing in this modern computerized world. People can use these search engines to gather information and to communicate together. Searching relevant information in the Internet is hard task. Usually, this searching process in the search engines can be done by using list of links available on the site. This link is nothing but a keyword; when choosing the keyword, the search engine can navigate through the web pages that are related with that keyword. Web Crawler is a basic entity or a program that facilitates the search engine to work efficiently on the World Wide Web. The main purpose of the crawler is to indexing the web pages that are visited by the user. Web Sense offers security solutions when it is implemented as a software applications or cloud-based service operating at the transport layer. This security solution is used by many business and government organizations to protect their networks from cybercrime, malware, as well as prevent the users from viewing inappropriate sites.In our research paper, we implement a methodology in order to validate the user, while browsing on the Internet. This can be carried out by tracking the user activities on the Internet through Web Crawler and validate the action through Web Sense, for security reasons and to avoid illegal access. Thus this paper provides a better solution to restrict the number of user browsing and to prohibit the user from accessing irrelevant sites.*

Keywords*: Cloud-based Service, Communication, Cybercrime, Information Technology, Internet, Malware, Search Engine, Web Crawler, Web Sense, World Wide Web.*

I. Introduction

As the evolution of the Internet is rapidly grown, the way of people communicate and work has also been dramatically changing. Most of the task in our daily life has been carried out through Internet, such as Banking, Billing, Shopping, and so on. The World Wide Web, WWW is a system of interlinked hypertext documents accessed via the Internet. WWW, also called W3 is growing at a fast rate, hence the user has to use this to view the information they desires. The Internet is a vast area which contains various categories of websites that are used by the people. People use the Internet for various purposes such as Communicating with each other; Search and gather required Information; and so on. For a particular category, there exists many websites, which are maintained and listed by a search engine. Search engines use a special component called **Crawling** to index the websites.

Web Crawling is a search engine, which is also termed as a web spider, web scutter, robot. The purpose of the web crawling is to gather the web pages, in order to index them and support a search engine. The main objective of this web crawler is to quickly and efficiently gather as many useful web pages as possible using the hyperlinks. In short, the Web Crawler is termed basically as a program that retrieves and stores pages in a repository.

Web Crawling is a system of downloading bulk of web pages using the set of hyperlinks. Web Crawlers can be used for variety of purposes. Of those, it is most prominently used by the search engine to index the web pages and to allow the users to access their queries.

The queries can be searched by the user through set of keywords. By submitting the keywords, the search engine search for the web pages related to that keyword, from the repository. These keywords are used for some purposes by the search engine.

In previous days, Web content was static and predictable. But now-a-days, these web content is constantly changing with end users, evenly the "trusted" sites. These web sites are most vulnerable to attack. **Web Sense** is another important technique handled in this paper to provide security solutions for various problems faced through the Internet. Web sense understands both the content and the context of the Internet, and develops its intelligence into all its security solutions.

Websense may be implemented as a software application, computer appliance or cloud-based service operating at the transport layer as a transparent proxy, or at the application layer as a web proxy. In each scenario, the effect is that it can inspect network traffic to or from the internet for a targeted group of people.

Websense allows the administrators to block the access on websites based on the categories. It contains a lists of sites that may be blocked, either permanently or at a particular period of times. This web sense can provide the security by watching the sites continuously with certain criteria provided in the software. The criteria may be usually of keywords used in the Internet to search for a particular site.

For example, while watching cricket on the Internet, the keywords mainly used are Score, Run, Wicket, Six, Boundary, Sachin, and Match. By using these keywords, we can collect required information about the cricket on the Internet.

In some companies, watching cricket news inside the campus is strictly prohibited due to some reasons. In such case, they implement the software with Web Sense to block the access to that site. This can be carried out by inputting the software with the keywords and URL related to the cricket. When the user tries to access the site containing the keywords, websense can block the site from access.

Mainly, this Web Sense can be mostly used in an organization such as business, IT Company, Schools, Colleges, and so on; in order to avoid the people or employee to access the unwanted sites on the Internet. Also, it can be used to protect the networks from cybercrime, malware and data theft, as well as prevent users from viewing inappropriate content and block the employees from browsing non-business related websites.

In IT Companies or other private sectors, this can used to prevent access to sites known to be infected with malicious content, it can prevent malicious programs from connecting to outside sites, and can limit the amount of bandwidth used by individual computers in a network. Thus Websense Web Security includes web protection services, which continuously monitors the organization's web sites, categories and associated URLs for malicious activity to prevent them from being used in fraudulent attacks.

In this research paper, we have to implement the Web Crawling with Web Sense in order to monitors the user activity on the Internet and to prevent the user from access on particular sites. While blocking the access on the site using keywords, there may be chance to occur some problems due to existence of the same keyword on different categories of sites. One such problem is that if an organization wants to block the site containing details about cricket, web sense uses the keywords to search the required sites and block the site containing the keyword. But in such case, news site containing the keyword 'Sachin' is also been blocked. This can be avoided by properly managing the appropriate site by the web sense. In our paper, we develop a methodology in order to tackle this problem by maintaining a browser history and the configuration table to block the user access on unwanted sites.

Our paper will be developed by analyzing some of the relevant papers regarding this Web Sense and Web Crawling. Also our proposed methodology contains algorithm and it can be experimentally verified. These are all discussed in the following sections.

II. Related Work

Anthoniraj et al, in paper [1] described that Crawlers are basic entity that makes search engine to work efficiently in World Wide Web. Semantic Concept is implied into the search engine to provide precise and constricted search results which is required by end users of Internet. Search engine could be enhanced in searching mechanism through semantic Lexical Database such as WordNet, ConceptNet, YAGO, etc; Search results would be retrieved from Lexical and Semantic Knowledge Base (KB) by applying word sense and metadata technique based on the user query. The Uniform Resource Locater (URL) could be added and updated by the user to Semantic knowledge base so that crawlers can easily extract meta data and text which is available in specified web page. The proposed methodology enables web crawler to extract all meta tags and metadata from the web page which are stored in Semantic KB, hence search results are expected to be more significant and effective.

Navdeep et al, in paper [2] stated that search engine transfers the web data from one place to another. They work on client server architecture where the central server manages all the information. A web crawler is a program that extracts the information over the web and sends it to the search engine for further processing. It is found that maximum traffic (approximately 40.1%) is due to the web crawler. The proposed scheme shown how web crawler can reduce the traffic using Dynamic web page and HTTP GET request using asp.net.

To reduce the web crawler traffic many researchers has completed their research in following areas: In this author used dynamic web pages with HTTP Get request with last visit parameter [3]. One approach is the use of active network to reduce unnecessary crawler traffic [4]. The author proposed an approach which uses the bandwidth control system in order to reduce the web crawler traffic over the internet [5]. One is to place the mobile crawler at web server. Crawler check updates in web site and send them to the search engine for indexing [6].

Deepika et al, in paper [7], the large size and the dynamic nature of the Web increase the need for updating Web based information retrieval systems. Crawlers facilitate the process by following the hyperlinks in Web pages to automatically download a partial snapshot of the Web. While some systems rely on crawlers that

exhaustively crawl the Web, others focus on topic specific collections. In present paper the various types of crawlers are discussed. The paper also discussed several web crawler design issues along with their solutions.

Vikas et al, in paper [8] described as the size of the Web grows exponentially, crawling the web using parallel crawlers poses certain drawbacks such as generation of large amount of redundant data and wastage of network bandwidth due to transmission of such useless data. Thus to overcome these inherent bottlenecks with traditional crawling techniques they have proposed the design of a parallel migrating web crawler. They first presented detailed requirements followed by the architecture of a crawler.

Sathish et al, in paper [9] stated that in a large distributed system like the Web, users find resources by following hypertext links from one document to another. When the system is small and its resources share the same fundamental purpose, users can find resources of interest with relative ease. However, with the Web now encompassing millions of sites with many different purposes, navigation is difficult. WebCrawler, the Web's first comprehensive full-text search engine, is a tool that assists users in their Web navigation by automating the task of link traversal, creating a searchable index of the web, and fulfilling searchers' queries from the index. Conceptually, WebCrawler is a node in the Web graph that contains links to many sites on the net, shortening the path between users and their destinations.

Al-Masri et al, in paper [10], addresses issues relating to the efficient access and discovery of Web services across multiple UDDI business registries (UBRs). The ability to explore Web services across multiple UBRs is becoming a challenge particularly as size and magnitude of these registries increase. As Web services proliferate, finding an appropriate Web service across one or more service registries using existing registry APIs (i.e. UDDI APIs) raises a number of concerns such as performance, efficiency, end-to-end reliability, and most importantly quality of returned results. Clients do not have to endlessly search accessible UBRs for finding appropriate Web services particularly when operating via mobile devices. Finding relevant Web services should be time effective and highly productive. In an attempt to enhance the efficiency of searching for businesses and Web services across multiple UBRs, we propose a novel exploration engine, the Web service crawler engine (WSCE). WSCE is capable of crawling multiple UBRs, and enables for the establishment of a centralized Web services' repository which can be used for large-scale discovery of Web services. The paper presented experimental validation, results, and analysis of the presented ideas.

III. Proposed Methodology

3.1 Proposed Method

The aim of the paper is to develop a methodology to provide security on the Internet, through Web Crawling and Web Sense. Now-a-days, Web Security is essential, since many illegal actions and malicious attacks have been taken place in the Internet. In order to avoid these actions, we have to propose a new methodology with Web Sense.

Web Crawler is nothing but a search engine, which manages a set of hyperlinks in the Internet. Upon selecting a link, the crawler can navigate the page from one to another based upon the query. The main purpose of the web crawling is to index the web pages. It keeps track of all the pages that the user visit, for later processing by a search engine which indexes the downloaded pages.

Web Sense is a software for providing security on the web and it acts like a sensor. It senses the user activities by keeping track of the pages they navigate, with the help of the keywords. It checks whether the user browse the page containing the keyword or not. Thus, by using this web sense, we can able to keep track of the user browsing history and to block the access on a particular site.

In most of the companies, the administrator provides some criteria that must be followed by the employee within the concern. The main criterion is to limit the access on Internet within the concern. That is, prohibit the user to access unwanted sites on the Internet. For example, in some companies, the employee is prohibiting to access their mail account. This can be inputted into the Web Sense software. Thus, when the user tries to access the mail account as "mail.yahoo.com", they cannot be able to access the site. The site cannot be opened. (i.e)the site is blocked by the web sense. Thus the web sense provides security to the web by monitors the client activities.

In some case, it goes wrong by getting confused with the keywords on the site. For example, in an organization such as schools or colleges, the students are prohibited to access on the cricket site by inputting the keywords related with the cricket on the Web Sense Software. When the students tries to open the cricket website or site containing the keywords related to the cricket, the software blocks the user access. If the student wants to know the information about 'Sachin', he tries to access the site by providing the keyword 'Sachin'. But the web sense software blocks the access to that site, since it contains the keyword 'Sachin' which is related to the cricket. Thus, it provides wrong result in some situation.

In order to tackle this problem, we have to propose a new methodology in our paper which is described below:

Instead of using the keyword to block the site, we have to maintain a LOG about the user browser history. This Log table contains details as follows:

Table 1: User Browser History Log

User Name	URL Visited	No. of Files/ Content Downloaded/ Viewed	Page Size	Date	Time	Status

This log table maintains the details about the user browsing details along with date and time; the name of the user; the URL visited by the user; what are the contents viewed by the user or downloaded by the user; the size of the page viewed or downloaded. When the user enters into the Internet, the Web Sense software collects all the necessary information about the action carried out by the user and the entry is added into the log table. The status of the user may be *active*, until the user gets blocked by the web sense software.

The organization also programmed into the software with the ***Configuration Table***. This configuration table contains the details about the limitation of the user browsing. (i.e)what is the maximum limit that a particular user can use the Internet per day. The structure of the configuration table is shown below:

Table 2: Configuration Table

User Level	Maximum Size / day	Maximum no. of times allowable to access the Internet/ day

This configuration table contains the details such as the level of the user in the company; what is the upper limit of the file size that the user can viewed or downloaded per day; and how many times that the user is allowed to access the Internet per day. The level of the user is the position of the user in that concern. Each user is allocated with these criteria to browse on the Internet.

When the user enters into the site, the web sense monitors the user activity and makes a entry into the Log table. Before making the entry, it compares the user level with the configuration table. If the user is a new one or the user not crossing the upper limit set, then the user is allowed to access the site and the entry is made into the log table. Otherwise, the user is blocked from accessing the site.

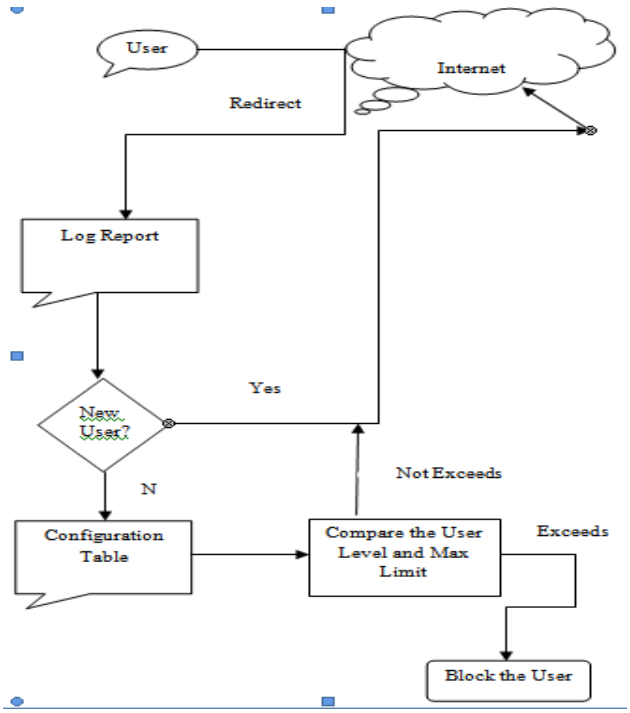

Figure 1: Proposed Architecture

Thus, our proposed methodology uses the web sense to reduce the usage of the Internet in an official or important concern, rather than block the site. That is, the user can view all sites and gather information, but with a limitation. This limitation is set in a configuration table. When the user crosses the limit, they are to be blocked to access the site further on that particular day. Each day the log table gets cleared to get the new entry. If the user is allowed to access for 1GB data for a day, then when the data limit exceeds, he/she can be blocked to view the site or download the content. Likely, when the user often visits the Internet frequently, they can also be blocked from access. Thus, based on the frequently accessed sites and the size of the content used by the user, the software blocks the user from further processing.

The main advantages of our proposed paper are not to block the site often; and only to reduce the usage of the Internet by restricts the user browsing.

3.2 Algorithm

```
Start
Administrator Builts a Configuration Table based on the User Level
Config_Tab={User_Level,Max_Size/day,
                    Max_no_of_times_used}
Maintain the Log Table
Log_Tab={Username,URL,No_of_Files_viewed/Downloaded        Page_Size, Date, Time, Status}
If user starts Internet then
Analysis the User Level
If Date.Now exists Log_Tab.Date and
   Log_Tab.Status=”Active” then
   Count = n(Username)
   If count >= Config_Tab.Max_no_of_times_used then
         Block the User
         Log_Tab.Status = “Blocked”
   Else
         Allow the user to access the site
         If the user starts Download then
            Sum = sum (Log_Tab.Username.Page_Size)
            If sum >=Config_Tab.User_Level.Max_Size/day then
Message “Can’t Download this Page”
 End if
 Start Download
Make entry into the Log_Tab
End if
End if
End if
End
```

3.2.1 Algorithm Explanation

The administrator builds a configuration table based on the user level with the information such as Maximum size downloaded per day and maximum number of times used per day. Then the log table is cleared.

When the user starts accessing the Internet, first the user level is analyzed and the information about the user is gathered from the configuration table. Then the software checks the log table for the user entry on that current date. If the entry exists, then the status of the user is verified. If the status is set as “Active”, then it counts the number of times that the user accessing the site and then compares it with the configuration table. If the number of times gets exceeds, then the user can be blocked by the software and the status is updated in the log table. Otherwise, the user can be allowed to access the site.

While accessing the site, if the user starts downloading the page, then the allowable size for the user can be verified from the configuration table. If the maximum size reached, then the user can’t able to download the page. Otherwise, he/ she can download the page and so on. Thus, by analyzing the user activities and with the criteria build by the administrator, our proposed algorithm performs well to control the user browsing and to prohibit the user from unwanted access on the site.

IV. Experimental Data

Our proposed methodology can be experimentally verified by implementing our methodology in various organizations and the results have been taken. One of our experiments has been carried out in an IT field, where the configuration file has been set up for different level of employee and their activities are

monitored through Log table. Our proposed method performs well compared to the existing method and it provides better results.

V. Conclusions

Crawler and Web Sense are most important component of a search engine from page downloading point of view. Due to explosive size of the web, many malicious activities can be taken place. Nobody can able to stop those activities from all sources of Internet. But we can able to control the malicious activities within a concern by means of Web Sense. This can be carried out in our paper and the methodology is experimentally verified. The experimental result also shows that our proposed methodology performs well. Thus we successfully implemented a methodology to reduce the malicious activities within a concern and to restrict the number of browsing carried out by the user.

References

[1]. Anthoniraj Amalanathan, Senthilnathan Muthukumaravel, "Semantic Web Crawler Based on Lexical Database", IOSR Journal of Engineering, Vol. 2(4), Apr. 2012.
[2]. Chetna, Harpal Tanwar, Navdeep Bohra, "An Approach to Reduce Web Crawler Traffic Using ASP.Net", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-3, July 2012.
[3]. Articles about Web Crawlers available at - <<http://en.wikipedia.org/~/~/# Examples_ of_ Web_ Crawlers.
[4]. Yuan, X.M. and J. Harms, "An efficient scheme to remove crawler traffic from the internet", Proceedings of the 11th International Conference on Computer Communications and Networks,14- 16, IEEE CS Press, (pp: 90-95), Oct 2002.
[5]. Ikada, Satoshi, "Bandwidth Control System and method capable of reducing traffic congestion on content servers" Dec 2008.
[6]. Bal. S and Nath. R, "Filtering the Web pages that are the not modified at remote site, without downloading using mobile crawler". Information Technology journal 9(2), ISSN 1812-5638, Asian Network for sciencetific information, (pp: 376-380), 2010.
[7]. Deepika, Dr. Ashutosh Dixit, "Web Crawler Design Issues: A Review", IJMIE, Vol 2, Issue 8, ISSN: 2249-0559, August 2012.
[8]. Abhinna Agarwal, Durgesh Singh, Anubhav Kedia, Akash Pandey, Vikas Goel, "Design of a Parallel Migrating Web Crawler", Volume 2, Issue 4, ISSN: 2277 128X, April 2012.
[9]. Dhiraj Khurana, Satish Kumar, "Web Crawler: A Review", IJCSMS International Journal of Computer Science & Management Studies, Vol. 12, Issue 01, January 2012.
[10]. Al-Masri E, Mahmoud Qusay H., "WSCE: A Crawler Engine for Large-Scale Discovery of Web Services", ICWS Conference Publications, 2007.