RESEARCH ARTICLE OPEN ACCESS

A Review on Web Content Filtering, Its Technique and Prospects

Ankur Baishya, Dr. Sangeeta Kakoty
Department Of Computer Science and Engineering
Assam Down Town University, Guwahati
Assam - India

ABSTRACT

Technological advancement increases the uses of Internet access and also web access. As the web access is increases, the vulnerability point in an organization, institution in any places is also increases which may affect the productivity and performance level of the organization. Many times it seen that the ways some places use the Internet is very much objectionable for others. As a result the whole network may not be fully utilized. In order to screen out the unwanted information from the access, content filtering technique is developed. Till today, lots of different filtering techniques are developed and developing by using different approaches, knowledge and technologies, but not fully able to protect the web content. This paper provides an inclusive survey of all major approaches, tasks and processes involved in the web content filtering and trying to find out a better technique to solve the above mentioned problem for the web access.

Keywords :— Web access, Content Filtering, Filtering Technique, Web protection.

I. INTRODUCTION

In our day to day life internet is widely used in almost everywhere such as home, corporate sector, academic environment etc. But it has seen that the network is not properly utilized. Most of the time people are seen to be busy in surfing other unwanted sites with unwanted information, especially the young aged students. Because of this web content filtering become more important in our society so that we can restrict few sites or information up to some extent. It has high demand now-a-days and it become an important research areas in the internet technology. The main reason is to proper utilization of network and also to overcome from the network junk. According to R.Ramli there is scant need to justify the requirement of the Internet and the World Wide Web (www) in today's education and corporate setup. The impact of the information revolution is double edged.

Now a day internet has become a crucial part of our modern life. It making our life smooth and easy, on the other hand people are concerned about its dark side especially to younger generations [1]. If these problems will increase day by day then it will be a big challenge for us to control these types of issues. At present scenario, there is no such type of automatic blocking controlling application developed and therefore results are reduced productivity, misuse of network resources and legal issues etc. In order to overcome these types of issues, we propose a new model with integration of Expert Filtering technology in content base domain area and an effective result towards our coming generation.

II. WEB CONTENT FILTERING

Web content is the visual and textual content which is use on a websites. It may include text, sounds, videos, animations and images. There are basically two types of web content. One is Textual web content and another is Multimedia web content. The textual web contents are those kinds of text or documents which are written for the web or internet. The other type of Web content is multimedia web content that includes images, video, animation sound etc. along with the content.

Due to the growth of traffic, the lack of central management on the Internet, and to prevent people from seeing offensive or inappropriate material on the internet. Some software's are use to block the certain sites. This kind of activities has been declared "Web Filters". Web Filters are designed to improve the security and productivity of a network, but as with anything else, it must be implemented correctly to work properly [2].

Content filtering is a technique using in a program or software to prevent certain types of things which may be harmful to access. In present scenario basically two common items to be filter, which are websites and emails. Generally content filtering works by matching strings of characters. We can say that it is a part of Internet firewalls. Generally content filtering is use in security purpose. It helps to decide which contents are acceptable for access and which are not. According to V.K.T.Karthikeyan web content Filtering is a firewall to block that kind of websites or web pages which are harmful or objectionable. Web Content filtering offers us various services, those are Web filtering, the screening of Web sites or pages and e-mail filtering. Some popular filtering techniques are Browser based filters, E-mail filters, 3 Client-side filters, Content-limited (or filtered) ISPs, Network-based filtering, Search-engine filtering etc [3].

Content Filtering Systems on web pages is an important research area primarily due to following reasons: It protects users i.e. minor ages people from unwanted content; the resources on the network can be saved from unwanted usage like playing network games, watching adult video in an office network etc. There exist many processes in Content Filtering

Systems. Basically these are: Rating Systems, Black List / White List, Keyword blocking etc. In the Rating System users are asked to rate a web site for its content. This rating could be used as a tool for filtering purposes. The black list / white list maintain a set of URLs manually prepared for filtering. Keywords are added with categories, and then used to protect against sites that had not explicitly added to the Master Database or defined as a custom URL [4].

According to V S Kharad and S. S. Kulkarni There are three major content filtering approaches. Those are Keyword analysis, packet analysis and URL analysis. In this case of Keyword analysis approach, entered keyword by the users are compared with the keyword present in the database, if match found then site containing that keyword is blocked. In this case of Packet analysis approach, IP address of packet is analyzed so that from where content has come from can be examined, because every packet is come from a particular IP address. In URL analysis process, a particular URL address of website is blocked because every site have a particular URL address , but a developer rearrange websites regularly, where top level address may not change but low level address may change quite frequently [5].

III. FILTERING TECHNIQUES AND TOOLS

In this section we discuss about some filtering technique and tools which are use in different purpose and different environment. Now a day data mining techniques are broadly use in the content filtering. According to Uri Hanani, Bracha Shapira Web mining is a technique in data mining that automatically retrieves extracts and analyzes the information from web. Wu et al and Yang,(2006) discuss about several issues like Over Filtering, Automated Data Cleaning, over sampling of data and Under Fitting techniques which are addressed in data mining techniques [6]. According to Berry Michael W Text Mining is a technique to introduce by computer system of new, formerly unknown information, by automatically takes out information from different resources [7]. Ou Wu and Weiming Hu have perform a work to filtering sensitive text by combining statistics and semantics property and examined them to build the cellular neural network like word net. This can help to extract right clues, text and helps to avoiding blocking normal text [8]. G. Aghila and K.S.Kuppusamy have developed a client side filtering model. This model can block the content of the whole website. This website is divided into some portion and blocks those portions which contain unrelated data. This particular experiment gives us 88% accuracy [9]. Jason Qin, Jianping Zhang and Quiuming Yan developed an application of web filtering which is based on novel URL. Machine learning algorithms and Maximum entropy algorithm are used In this technique to break the URL in n-grams. This method is useful only for text based web sites [10].

The basic Content filtering tools are k9, Dans Guardian, Open DNS, Squid Guard/Squid and Kinder gate parental control. K9 is a desktop solution application. After installing the software it scans all the requests we make towards the filters we have specified. But there have some restriction of working from a static database. In order to overcome this problem k9 introduce the Dynamic Real Time Rating System. Dynamic Real Time Rating System accesses the whole Content of the Websites and bans the particular websites if the website fall into the filter categories. Dans Guardian runs on Linux, NetBSD, Mac-OS, HP-UX, Solaris and FreeBSD. It allows us to Filters unwanted ad's in our network, Blocking images, Block the extension of the downloading files etc. Open DNS is a application to set-up & manage the content-filtering server. It replaces our current DNS server. It filters every connection which is sending out from the source as machine. If anybody is on main machine or connecting into the wireless via desktop or laptop, every single thing will be filtered by Open DNS. Squid Guard / Squid is Similar to Dans Guardian. We can install it on Linux. It is a standalone filtering tool to connect the proxy. It is a personal internet security solution. We can say that It is real time content filtering system. It uses HTTP traffic filtering, blocking of unwanted web sites, deep content inspection and pages [3] [11].

IV. RELATED WORKS

Po-Ching Lin,Ying-Dar Lin and Yuan-Cheng Lai developed an algorithm namely “An Early Decision Algorithm” to filtering the web content from a particular resources. This algorithm speeds up the web filtering techniques in the web content either to pass or block the web pages. This technique is used the testing samples of DansGuardian. This work addresses the problem of long delay from text classification algorithms to perform the runtime content analysis of web content. But, it did not combine with more keywords and maintains the URL list [12].

Neha Gupta and Saba Hilal have done a work which is very helpful for teachers, parents and kids to search the web in secure way and make delightful communion. This works makes children to focus the websites by connecting the interest into educative and informative side by not blocking or filtering only by connecting their WebPages. There is no need for blocking, keep kids interest on educational side, safety form security threats, mainly for play psychologically with their child [13].

Akebo Yamakami and Tiago A.Almeida presented a relative study of linear support vector machines, Naive Bayes classifiers, Mathews Correlation Coefficient prediction. Then, it manages the test on big data set as TREC05, TREC06, TREC07. They compare the perspective with other open source anti spam filters such as Bogo filter, Spamong others [14].

V.K.T. karthikeyan and M .Thangaraj developed "KT-GRANT algorithm". This algorithm first examines the Web content to a definite category first, after that makes the filtering decision, either to ban or allow the content. The filtering work of this algorithm present a simple, but effective in the form of observation that the filtering decision can be made scanning through the entire content of the web page, as soon as the content can be analyzed into a certain category[15].

Liming Chena, Mohamed Hammanai and Youssef Chachir have developed a web based adult content filtering and detection system, Name is "web grant" system. With the help of this technique we can extract relevant data from web. With the help of data mining technique it extract images, text, URL names and analyze them [16].

Willy Susilo, Rongbo Du and Reihaneh Safavi Naini developed an algorithm which is used image and text classification technique. This algorithm is used to allow or block those kinds of web pages which contain prohibited contents or images [17].

Zhifang Liao, Hui Li and Fei cai developed an algorithm which is used to block unwanted messages of a social media websites. With the help of this algorithm user have to control on his wall of his social networking sites. It also able to increase the efficiency of the filtering techniques. Machine learning technology is used in this algorithm [18].

Mingliang Zhu, Ou Wu, Weiming Hu and Zhouyao Chen developed an algorithm namely novel base web page filtering. This algorithm is used to filter objectionable images and text from web sites [19].

Sarifullah Khan ,Sadaf Khurshid and Shariq Bashir have developed an smart filtering technique to classify the text, in this technique they using feature engineering method and sentiment analysis of text. In order to block the unwanted web pages Novel content filtering technique is used. With the help of machine learning algorithm Text classification is done to classify the positive and negative classes [20].

Ammar Almomani , Eman Almomani, B.B.Gupta and Samer Atawneh A Meulenberg have developed a technique to avert phishing email attacks. This is basically a client based filtering technique. Supervised learning algorithm is used in this technique. With the help of this technique we can able to identify new email attacks [21].

V. CONCLUSION

This paper presents the detail description of the web content filtering and also reviewed different types of techniques and tools used in filtering and summarized them. In this paper we discuss about Keyword analysis, packet analysis, URL analysis, Rating Systems, Black Listing / White Listing, Keyword blocking etc. we also reviewed novel filters technique, client side filtering, web grant system, KT-GRANT algorithm, Early Decision Algorithm etc. From the study we observed that in almost all approaches there have some limitations, some have not fully filtered as per requirement. Therefore, a perfect, compatible to all platform and cost effective approach is urgently required in this area.

REFERENCES

[1] R.Ramli, "The Internet, its Social and Ethical Problem to the Young and How Curriculum Can Address the Issue", International Journal of Social, Behavioral, Educational, Economic, Business and Industrial Engineering, Vol. 5, No.11, p.1612, (2005).

[2] Robert Alvey, "The Art of Web Filtering", GSEC Practical v1.4b, (2004).

[3] V.K.T.Karthikeyan, "Web Content Filtering Techniques: A Survey", International Journal of Computer Science & Engineering Technology (IJCSET) ISSN: 2229-3345 Vol. 5, No 03, (2014).

[4] K.S.Kuppusamy and G.Aghila, "A personalized web page content filtering model based on segmentation", International Journal of Information Sciences and Techniques (IJIST) Vol.2, No.1, (2012).

[5] V S Kharad and S. S. Kulkarni, "International Journal of Advanced Research in Computer Science and Software Engineering" Volume 5, Issue 4, p. 248, (2015).

[6] Uri Hanani, Bracha Shapira and Peretzshoval, "Information filtering: overview of issues, research and systems", Department of Information System Engineering, PP. 203-259, (2001).

[7] Berry Michael W, "Automatic Discovery of Similar Words", in "Survey of Text Mining: Clustering, Classification and Retrieval", Springer Verlag, New York, PP. 24-43, (2004).

[8] Wu and Weiming Hu, "Web sensitive text filtering by combing semantics and statistics," Preceding of NLP-KE IEEE, (2005).

[9] K.S.Kuppusamy and G. Aghila, "A personalised web page content filtering model based on segmentation", International journal of information sciences and techniques Vol.2,No1, (2012).

[10] Jianping Zhang, Jason Qin and Quiuming Yan,"Role of URL"S in objectionable web content cauterization",Procedding of international conference on web intelligence IEEE ,(2006).

[11] Sangita S. Modi and Dr. Sudhir B. Jagtap, "Web Content Filtration Using Different Web Mining Techniques in Educational System: An Overview", International Journal of Advanced Research in Computer Science and Software Engineering 7(3), pp. 135-139, (2017).

[12] Ying-Dar Lin, Po-Ching Lin, Yuan-Cheng Lai. "An Early Decision Algorithm to Accelerate Web Content Filtering", IEICE TRANS. INF. & SYST., VOL.E91–D.

[13] Saba Hilal and Neha Gupta, "An Algorithm to Filter & Redirect the Web Content for Kids", IJET13-05-01-024.

[14] Tiago A.Almeida and Akebo Yamakami, "Content Based Spam Filtering", Neural Networks (IJCNN), The 2010 International Joint Conference on 18-23 July, (2010).

[15] M. Thangaraj and V.K.T.Karthikeyan, "KT-grand: an algorithm for web content filtering", international journal of advance research in computer science and management studies", Vol2, Issue 9, (2014).

[16] Mohamed Hammanai, Youssef chachir, Liming chena "Web Guard: Web based adult content detection and filtering system", Procedding of international conference on web intelligence, IEEE, (2003).

[17] Rongbo Du, Reihaneh Safavi Naini and Willy Susilo, "Web filtering using text classification", IEEE, (2003).

[18] Hui Li,Fei cai and Zhifang Liao, "Content based filtering recommendation algorithm using Hmm", International conference on computational and information sciences, IEEE,(2012).

[19] Zhouyao Chen, Ou Wu, Mingliang Zhu and Weiming Hu, "A novel web page filtering system by combining texts and images", Procedding of international conference on web intelligence, IEEE, (2006).

[20] Sadaf Khurshid, sarifullah Khan and Shariq Bashir, "Text-based intelligent content filtering on social platforms", International frontiers Of Information Technology l, IEEE, (2014).

[21] Ammar Almomani, B.B.Gupta, Samer Atawneh A Meulenberg, and Eman Almomani, "A survey of phishing email filtering techniques",IEEE communication surveys and tutorials Vol 15, No. 4, (2013).