Volume 7, Issue 3, March 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Web Content Filtration Using Different Web Mining Techniques in Educational System: An Overview

[1]Sangita S. Modi*, [1]Dr. Sudhir B. Jagtap
[1]Research Scholar, [2]Professor and Principal
[1, 2]Research Centre in Computational Science, Swami Vivekanand Mahavidyalaya, Udgir, Dt:- Latur, S.R.T.M. University, Nanded, Maharashtra, India

DOI: 10.23956/ijarcsse/V7I3/0122

Abstract— Internet is widely used as ICT tool in today's education. The web content filtering is essential tool used to filter offensive, unwanted web content from web pages. Internet is widely used in educational organisations as an ICT tool. To protect our tech savvy student from accessing adult sites, offensive, unwanted data in absence of teachers and parents in school as well as home. A strong filter is required to prevent this type of unsolicited activity. In this paper we review the existing filter tools and their techniques.

Keywords— Data mining, Web mining, ICT, web content filter, tech savvy, offensive,.

I. INTRODUCTION

A website filter is a network application used for website control and or to manage traffic. Website filters are used as a tool and security features to block network traffic according to a user or network preferences. Website filters are built into devices or software including router, switches, firewalls, anti-spyware software, and browsers. Web filters are the tools, designed using web mining technique. Web mining is an application of data mining.

ICT is an umbrella term that includes any communication device or application, encompassing: radio, television, cellular phones, computers, and internet etc. To improve quality of teaching and learning we are using Information communication technology (ICT) is used to enhance the learning process. While surfing on internet, web Filter is used to filter unwanted data, inappropriate content across the web, and at the same time allowing students to use rich educational sites to enhance their learning as well as knowledge. Technology has brought more computing to schools, and so that, it need to protect students from inappropriate content across the web. Instead of blocking off large portions of the Internet, many schools are utilizing customizable web filtering systems that allow to control over which sites are allowed or blocked. the Child Pornography Prevention Act of 1996 (CPPA), the Children's Online Privacy Protection Act of 1998 (COPPA) and the Children Internet protection Act (CIPA) requires that schools have to protect students from obscene or harmful online content in order to be eligible for discounts on internet access or internal connections through the Schools and Libraries.

II. THE PROBLEM IN EDUCATIONAL ORGANIZATION

Most educational sites are AdSense enabled for their revenue. Websites displays advertisements which contain all types of products. When any child accesses e-learning sites, automatically advertises or offensive contents are displayed in the form of image or videos on the web pages. This may divert the learner's attention to any e-commerce or offensive sites. This is one of the increasing problems for parents and teachers in today's world. The government is urging to educational institutes to use information communication technology (ICT) tools for teaching and learning process. So, parents and guardians require internet with parental control. There are purposely built filters that analyses each page for inappropriate content before deciding whether it's suitable for the user or not. If so, the request is blocked and the user is sent back to a safe location. Introducing these tools at home and educational organization would be the simple answer to prevent access to adult content. It allows parents to keep control over their child's web access.

Educational organization try to facilitate good infrastructure in respect to information technology with Wi-Fi enabled campus. Students are also able to access social, adult content sites at the class or at home. But without web filtering it is difficult to manage the login from improper content to the students and other related member. So web filtering is prime factor in ICT education.

III. TYPES OF WEB CONTENT FILTERING

1) *Server Side Filter*: In this content filtering software is installed on a central server which can monitor the security settings on all the other systems on the same network. The network administrator can apply the same filter rules for all computers which are connected to server.
2) *Content Limited ISP*: In this filter, the internet service provider has the authority to regulate the types of pages that can be not content any unwanted data for the users. It is used block malicious websites, monitors emails, chats and web traffic to prevent Denial of Service (DoS).

3) *Search Engine Filters*: Many search engines like yahoo, Google and Bing also offer content filtering options. They can block inappropriate content from being displayed in the search results.
4) *Client Side Filter*: In client side filtering, software is installed on computers that require content filtering. The admin can customize the list of blocked websites or specify guidelines according to which the content needs to be filtered. Client side filters are a good option for educational organization and small business. Filter toolkit is installed which is worked with firewall.
5) *E-mail filters*: It filters information contained in the mail headers such as sender and subject, and file attachments etc. to accept or reject the messages. It is a Network-based filtering: It is implemented at the transport layer as a transparent proxy, or at the application layer as a web proxy. Filtering software may include data loss prevention functionality to filter outbound as well as inbound information. All users are subject to the access policy defined by the institution. The filtering can be customized as per user or group user requirement.
6) *Social Networking filter*: It is used to filter some unwanted text from social networking sites. This filter is used for Facebook, twitter type of the sites where somebody posting offensive text or content.

IV. RESEARCH WORK CONDUCTED ON WEB CONTENT FILTERING

In this paper Sadaf Khurshid, Sarifullah Khan and Shariq Bashir, have done work on an intelligent filtering technique using sentiment analysis of text and feature engineering method to classify the text. Novel content filtering technique is used to block the unwanted web pages. Text classification is done with machine learning algorithm which classifies them in positive and negative classes. Navie bayes is a text classifier used to classify unknown samples. Another J48 decision tree classifier is used to find relationship between features and classes.[1]

In this paper Jianping Zhang, Jason Qin and Quiuming Yan, have done work on a novel URL based objectionable content categorization approach and application of web filtering. In this model Maximum entropy algorithm, machine learning algorithms are used to break URL in n-grams. They also used supervised learning algorithm to classify URL's. This method is useful only for text based web sites.[2]

In this paper Ammar Almomani, B.B.Gupta, Samer Atawneh A Meulenberg, and Eman Almomani have done work to prevent phishing email attacks. Various techniques are used to detect such type of phishing mails. This is client based filter works offline. In this technique supervised learning is used, through which it can detect new email attacks. Unsupervised learning is used which is faster but having low accuracy and hybrid learning which is time consuming and costly.[3]

In this paper Ou Wu and Weiming Hu have done work to filter sensitive text by combining semantics and statistics features and analyzed them to construct the CNN (cellular neural network) like word net. This can help to extract right clues, text and helps to avoiding blocking normal text. For classification support vector machine classification technique is used to find sensitive word. But in this paper not done any work on misspelled problem. [4]

In this paper Zhouyao Chen, Ou Wu, Mingliang Zhu and Weiming Hu have done work on a novel web page filtering. It is used to filter offensive text and images from web sites. In this method divide and concur, navie bayes, k-nearest neighbour algorithms are used.[5]

In this paper Hui Li,Fei cai and Zhifang Liao have done work on a filter which is used to block unwanted messages and allow user to have direct control on the message posted on the wall of online social networking sites. Inference algorithm is used to infer the new information from the filtering rules to increase the efficiency of the filtering process. Machine learning is used for content based filtering.[6]

In this paper Rongbo Du, Reihaneh Safavi Naini and Willy Susilo have done work on Web filtering which is used for text classification. The proposed algorithm is used to block or allow the web page which contains forbidden contents. It also identifies images on the pages with 'alt' attribute in image tag of HTML page. It also block those web pages which contain forbidden images.[7]

In this paper K.S.Kuppusamy and G. Aghila have done work on a model which is client side filter. This filter can block the content of whole website or page. The web page is divided into segment and blocks those segments which contain irrelevant data. After the experiment it gives 88% accuracy. Document object model is used in segment filter, which is applied on text, images, link in the web pages.[8]

In this paper M. Thangaraj and V.K.T.Karthikeyan have done work on KT grand web content filter algorithm. This algorithm is used for filtering web content in web pages. It analyses the content and then decided to block or allow the web page from access.[9]

In this paper Mohamed Hammanai, Youssef Chachir, Liming Chena, have done work on a filter 'Web Guard', which is web based adult content detection and filtering system. Web Guard' uses web crawler to extract relevant data from the web. It extract text, images, URL names and analyse them with data mining technique.[10]

V. WEB FILTER TOOLS

There are number of solutions are available for schools/colleges/institutions their web filtering needs.

1. *DansGuardian* (Cross Platform, Free): Dansguardian runs on Linux, FreeBSD, OpenBSD, NetBSD, Mac OS X, HP-UX, and Solaris. It is extremely configurable and allows you to do all sorts of things, like block all images, filter ads out across your entire home network, block files from being downloaded by extension type, and control the effects of the filters, whitelists, and more based on which computer on your network is doing the accessing. You can deploy different filters for different computers based on domain, user, and source IP.[11]

Figure 1 Dans Guardian[10]

2. Kinder gate parental control: this is home internet filtering security solution. It is real time content filtering based on mechanism of url filtering & morphological analysis filtering contextual advertises, safe search, black list, white list, and control of downloads. It complies the regulation of internet watch foundation (IWF) and Children Internet Protection Acts (CIPA). It uses HTTP traffic filtering, deep content inspection, blocking of unwanted sites, and pages.[12]

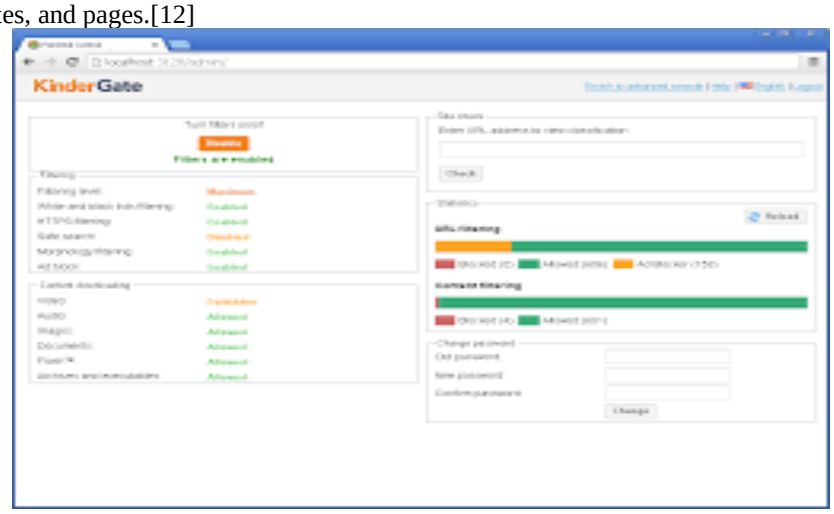

Figure 2 Kinder Gate Parent Control

3. K9 (Windows/Mac, Free): K9 is a desktop solution. Install the software and it checks all the internet requests you make against the filters you have specified. In an effort to overcome the limitations of working from a static database, K9 introduced Dynamic Real-Time Rating to actively access the content of websites and ban them if they fall into the filter categories you've selected. It main strong points are the division of filtered content into 60+ categories which is easy to manage. [13]

Figure 3 K9 Parent Control

4. OpenDNS (Cross Platform, Free): OpenDNS is a perfect solution for people who either lack the time or expertise to set up and administer a full-out content-filtering server. OpenDNS replaces your current DNS server and allowed to filter every connection coming out of your house if you change the DNS settings at the router level. No matter if someone is on main desktop or connecting into the wireless via laptop, everything will be filtered by OpenDNS.One can set custom filters to white list and black list specific sites and customize the range of filters.[14]

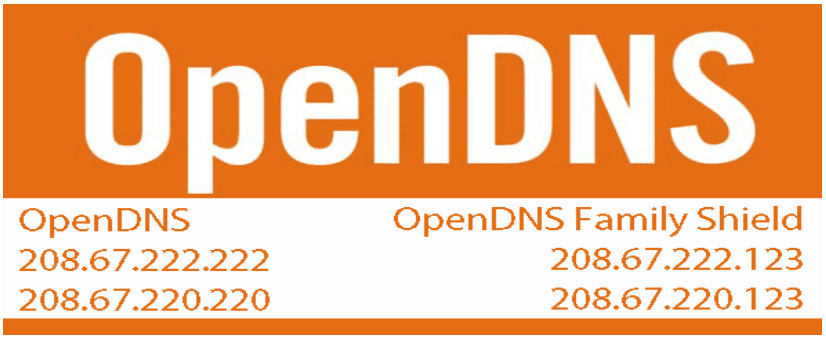

Figure 4 Open DNS

5. SquidGuard/Squid (Linux, Free): SquidGuard is natively aUNIX-environment only tool, and you can install it onto Linux, FreeBSD, and so forth. It is a standalone filtering tool you have to just connect into with a proxy and you are protected with filtering tools.[15]

Figure 5 Squid Guard

6. Securly is a cloud based subscription service that offers K-12 internet content filtering. It is designed in integration with Google Apps for education with chrome book. It is designed to prevent the problem of "over blocking" in schools and organization.[16]

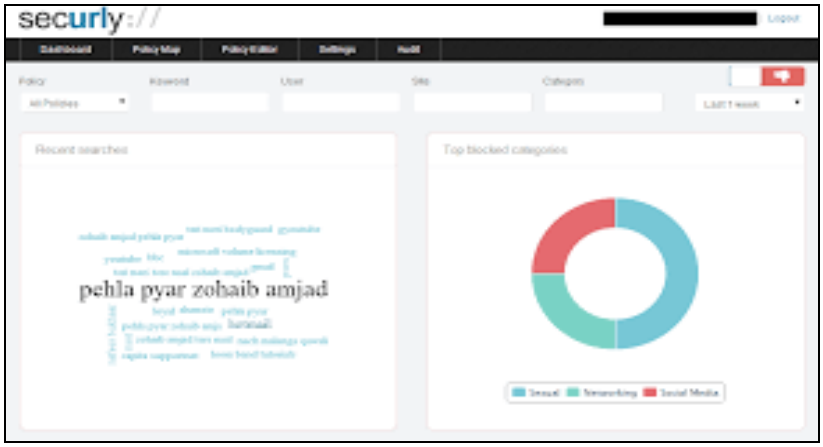

Figure 6 Securly

7. Netbox Blue has a database of over 6 Billion URLs that enable schools to easily allow or block access to web sites based on predefined categories. Teachers, IT staff or other school officials can add to or remove specific URLs if required.[17]

Figure 7 Netbox Blue

.

VI. CONCLUSION

In this paper we have reviewed novel filters, URL based filter, text filter, social networking sites filter, phishing email filter, and pornography content filter. We have also reviewed different types of techniques used in filtering and summarized them. These filters are useful to keep clean and enhanced learning experience with the help of ICT tools. Presently, so many web filters are built for educational organizations. But all are either offline or online tools. Here we suggest, using machine learning algorithm that, a filter should be designed in such a way that if a student try to access the irrelevant content knowingly and unknowingly then the filter should redirects the student to the site which he supposed to access. The desired filter should also inform administrator via email about the irrelevant access.

REFERENCES

[1] Sadaf Khurshid, sarifullah Khan and Shariq Bashir, "Text-based intelligent content filtering on social platforms", International frontiers Of Information Technologyl , IEEE , 2014.
[2] 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.
[3] 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.
[4] Ou Wu and Weiming Hu, "Web sensitive text filtering by combing semantics and statistics," Procedding of NLP-KE *IEEE 2005*.
[5] 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.
[6] Hui Li,Fei cai and Zhifang Liao, " Content based filtering recommendation algorithm using Hmm", International conference on computational and information sciences, IEEE 2012.
[7] Rongbo Du, Reihaneh Safavi Naini and Willy Susilo, "Web filtering using text classification", IEEE, 2003.
[8] 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
[9] M. Thangaraj and V.K.T.Karthikeyan, "KT-grand :an algorithm for web content filtering", international journal of advance research in computer science and manegment studies", Vol2,Issue 9,2014
[10] 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.
[11] http://dansguardian.org/
[12] https://www.entensys.com/products/kindergate-parental-control/overview
[13] http://www1.k9webprotection.com/
[14] https://www.opendns.com/home-internet-security/
[15] http://www.squidguard.org/
[16] https://www.securly.com/
[17] http://netboxblue.com/

