Browser Integrated Web Content Filtering Using Natural Language Processing

Sangita S. Modi*  
Research Scholar,  
Research Centre in Computational Science,  
Swami Vivekanand Mahavidyalaya, Udgir, Dt:- Latur.  
S.R.T.M. University, Nanded, Maharashtra, India.  
sangitasable1@gmail.com

Dr. Sudhir B. Jagtap  
Professors and Principal,  
Research Centre in Computational Science,  
Swami Vivekanand Mahavidyalaya, Udgir, Dt:- Latur.  
S.R.T.M. University, Nanded, Maharashtra,India.  
sudhir.jagtap7@gmail.com

Abstract: Today internet is one of the rapid growing ICT technologies which are used in every field like business, marketing, banking, entertainment, transporting and education etc. over the world. E-learning technology is playing a vital role in modern education which facilitates the student to learn anywhere, anytime, with their own speed and technology. So that, Internet filtering and blocking is very important for a couple of reasons. With the flooding of explicit and violent materials on the Web, educators and parents would like to block these offensive materials from their children. In this paper we proposed a browser integrated client side filter using natural language processing. The web content filtering is used on browser integrated extension which helps to keep children away from accessing offensive contents and encourages to safe e-learning. This filter performs web page text content mining to create keyword database of 2000 e-learning web sites which are collected by survey method. Extracted text from web page used as knowledge base to block the non learning sites in the educational organisation.

Keywords: web filter, ICT, e-learning. Natural language processing, web mining, browser, extension.

I. INTRODUCTION

The Internet has rapidly made a great impact on the world. With development of the www, Internet surfing has become an important daily activity for the people in general. Internet is used in communication, buying and selling of product online, travel booking, route tracking, online medical diagnostic center, entertainment, news sites, online education training centers, etc.

Education is one of the fields, which is growing swiftly. E-learning means electronic learning which is a novel concept that enables students to take their education anytime, anywhere, with their own speed and technology. These e-learning sites run different subject courses to improve knowledge of students by sitting at home or in school. Any student can avail this service on paid or on free basis. While using internet there is no restriction for the student in accessing the web content. Student may also try to access harmful data in the absence of parents at home or in the absence of teacher at school and colleges. So web filter plays an important role to block irrelevant websites. There are different web filters present in different forms in the computer system, which helps to monitor the student’s data accessing habits and try to block harmful sites. There are billions of websites that contain obscene and malicious information.

To access internet on computer system, users always used browsers like Google Chrome, Opera, Mozilla Firefox, Safari, Internet explorer etc. It is very difficult to restrict user from accessing unwanted sites at browser level. At browser level we can only use extension, plug-in, add-ons, and web application, which help to block unrelated web content. Different browser extension are used to block ads and adult contents, web data scraping, YouTube ads block etc. But no browser extension is found that restricts the student from browsing irrelevant sites. In our proposed work third party application i.e. browser extension is used to restrict the student in education organization from accessing non e-learning sites. Proposed Extension allows the student to access only e-learning sites.

II. E-LEARNING AND ICT TOOLS

In modern education system with the help of computer we can easily teach the concepts of science, mathematics, geography, economics, history, civics etc. The teacher communicates with students through computer tools which are known as information communication technology. To communicate with the student we use hardware as well as software tools. The hardware like keyboard, mouse, monitor, scanner, printer, speaker, microphones, projector, etc. are ICT tools used to enhance teaching and learning process. Also application like email, internet, MS Office, Adobe Reader, paintbrush, Geogebra, Selenium, etc. are used as ICT tools. The internet is one of the data warehouse through which student can access number of sites to acquire knowledge.
A. Browser

Through different web browsers student can access and view websites easily. In this process user sends request from client PC in the form of URL through browser. Web browser sends the request to that web server. Then the web server sends the response back to the client and displays it on the browser window.

B. Browser Extension

Extensions are browser based software programs that helps to modify the browsing experience better. They enable users to shape Chrome functionality and actions to meet the individual needs and preferences. Web technologies such as HTML, JavaScript, and CSS are used to develop extension.[1]

III. WEB FILTERS

A web filter is an application or utility which is specifically used to control the website traffic. Website filter plays a vital role in blocking the content as per the user and the network demand. Filters are built into hardware as well as software units like routers, switches, firewalls, anti-spyware software, and browsers. The network administrator is always configured to the web filter. Mostly website filters are used to block offensive web data, phishing mails, ads, viruses, pornography, fraud, etc. which are the most harmful things on the internet. Web filter are used in computer for security of digital assets. To accomplish this there are different types of filter like

1) Server Side Filter: This content filter installed on the main server and all the clients are connected to the server. The security setting of the network can be monitored by server. The network administrator applies the same filter rules to all rest of the computers which are centrally connected to the server.

2) Content Limited ISP: This filter is specially used to block websites which contains hateful data and monitors emails, chats, and web traffic to avoid Denial of service (DoS)

3) Search Engine Filters: Search engine filter contains web crawlers which block inappropriate websites being displayed in the search result. 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 this filter, 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 works 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 offensive text from social networking sites. This filter is used for Facebook, twitter etc. and other social networking sites where someone may post offensive text or content.

There are number of parental control tools available which are as follows.

1) The k9 web protection desktop application is used as content control software to block unwanted sites. It uses updated internet based database of blocking sites [2].

2) Qustodio is the Parental Control software offering a plethora of features to protect your children. This software helps to understand behavior of a child on web, supervise his online activity, restrict web usage by time setting, monitor social networks, and block unwanted sites and content on the internet [3].

3) OpenDNS is a preconfigured Family Shield to block adult content. It works on router level [4].

4) DansGuardian blocks all images, filters ads, and blocks files from being downloaded by extension types [5].

5) Kinder gate parental control is home internet filtering solution. It is a real time URL filter, which blocks advertisement, secures search, controls downloads and creates black list and white list. It uses HTTP traffic filtering, deep content inspection, blocking of unwanted sites, and pages [6].

6) Squid Guard is a standalone filtering tool which works at proxy level [7].

7) Securely is a cloud based K-12 internet content filtering tool. It is designed in combination with Google Apps for education with chrome book. It is designed to prevent the problem of "over blocking” in schools and organization [8].

These available filters are used to block adult, prone, offensive, violent and irritating content of web pages. But there are no filters are available for educational organization which allows opening only learning sites and block other non learning sites.

IV. RELATED WORK

According to Cohen Almagor the tools of client side filtering are familiar because they are straightforward to execute and provide guardians and parent a simple way to
offer a protective surrounding of internet. A similar personal use is a filter on client side installed on a home PC by a parent desired to secure child from improper content. Client side filtering is available in surroundings in which certain points of access in a LAN must be filtered [9]. Daugherty has stated that the client side filters have one major limitation. They do not stop junk electronic mail before it meets the user’s PC. Every user must acquire the liability for assuring that the filter is enabled and configured. Because the filter is not centralized it is difficult to use consistent parameters of configuration on entire systems of client. Instead every system must be designed separately. For this reason it is best to attack junk electronic mail issue at the server in addition to client side server [10]. According to the centextiguy the software is installed on PC that needs content filtering in client side filtering. The admin customize the blocked websites list or specify guidelines according to which the needs of content are filtered. Client side filters are a better choice for small businesses that have restricted number of workers [11]. The paper of Kuppusamy and Aghila has proposed work on a model which is a client side filter. This filter can block the whole page or website content. The web page is classified into partitions and blocks those partitions which comprise irrelevant data. After the experiment it provides 88 percent accuracy. The model of document object is used in segment filter which is used on images, text and connects in web pages [12]. Reimer et al has stated that organizations can lose control of their electronic mail easily or have to maintain and roll out solutions of client side content filtering. An organization wide archiving solution is meaningless if it is not feasible to access electronic mail content or an archival of client side solution has to be deployed for each separate user [13].

V. FRAMEWORK OF PROPOSED WORK

Website Mining is the method of extracting helpful information from the internet documents. Content information corresponds to the gathering of text, images, hyperlinks, audio, and video data from online webpage and processes that data to convey useful information to the users. Application of web text content mining like topic discovery, extracting association patterns, agglomeration of internet documents and classification of web content from website has been widely researched.

When student surfs on the web, knowingly or unknowingly he/she tries to accesses irrelevant sites. This is a challenge to prevent the student from accessing non educational sites on internet. There are billions of sites on internet which are non learning like social networking sites, sports, entertainments, ecommerce etc.

Our proposed framework is a novel chrome Browser extension. For this extension to create keyword knowledge base of learning sites, collected the list of non learning sites URLs Blocked(Ub1,Ub2…Ubn) and learning sites URLs Allowed(Ua1,Ua2,..Uan) by survey method. These steps are shown in algorithm-1. In the first phase of algorithm extraction of text contents is carried out from list of URLs Allowed sites. After the text extraction, it removes the stop words, duplicate words and finally creates database D’.

In next phase, from the list of URLs Blocked sites, extracts the text data and converted to word tokens T1,T2,…Tm.

Now remove all words of blocked sites i.e. T1,T2,…Tm from allowed site word database D’. This forms the unique keyword dataset of learning sites called as D, which is used by the extension as knowledge base.

The second algorithm is proposed to show actual working of extension to block all non learning sites. Using chrome browser, first of all accept the url ‘Ur’. Then extension extracts the text content from the requested web page. The extracted text data processed using natural language processing (NLP) to extract the all nouns N from web page.

After that it finds the percentage P of extracted nouns N by comparing it with knowledge base of learning dataset D. If percentage is more than threshold value ‘theta’ i.e. (P > theta) then extension will allow that web page to display on browser, otherwise it will block and redirect to google search engine page.

Algorithm1:-Dataset

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Step 1. Get the list of URL URLS_Allowed={Ua1,Ua2,..Uan}</td>
</tr>
<tr>
<td>2.</td>
<td>Step 2. Get the list of URL URLS_Blocked={Ub1,Ub2,...,Ubn}</td>
</tr>
<tr>
<td>3.</td>
<td>Step 3. For each U in URLS_Allowed</td>
</tr>
<tr>
<td>4.</td>
<td>Extract text TEXT_Allowed from U</td>
</tr>
<tr>
<td>5.</td>
<td>Step 4. Filter only alphabetic text i.e. TEXT_Allowed = Fa(TEXT_Allowed)</td>
</tr>
<tr>
<td>6.</td>
<td>Step 6. Tokenize TEXT_Allowed i.e. TOKENS_Allowed = {T1, T2,...,Tm}</td>
</tr>
<tr>
<td>7.</td>
<td>Step 7. Append TOKENS_Allowed to the keywords dataset D = {D1,D2,...,Di}</td>
</tr>
<tr>
<td>8.</td>
<td>Step 8. Remove duplicates D’ = unique(D’)</td>
</tr>
<tr>
<td>9.</td>
<td>Step 9. For each U in URLS_Blocked</td>
</tr>
<tr>
<td>10.</td>
<td>Extract text TEXT_Blocked from U</td>
</tr>
<tr>
<td>11.</td>
<td>Step 10. Filter only alphabetic text i.e. TEXT_Blocked = Fa(TEXT_Blocked)</td>
</tr>
<tr>
<td>12.</td>
<td>Step 12. Tokenize TEXT_Blocked i.e. TOKENS_Blocked = {T1, T2,...,Tm}</td>
</tr>
<tr>
<td>13.</td>
<td>Step 13. Remove if present; TOKENS_Blocked from the keywords dataset D”</td>
</tr>
<tr>
<td>14.</td>
<td>Step 14. Sort the tokens i.e. D = sort(D”)</td>
</tr>
<tr>
<td>15.</td>
<td>Step 15. Store the keywords dataset D</td>
</tr>
</tbody>
</table>
Algorithm 2: Keyword-based browser extension to block non-learning sites using NLP.

Step 1. Get the requested url 'Ur'
Step 2. Extract text T from the retrieved page with url 'Ur'
Step 3. Apply NLTK to extract nouns N = \{N1, N2, ..., Nn\}
Step 4. Find percentage of these nouns P = n(N Intersection D)/n(N) that are present in the dataset D = \{D1, D2, ..., Dn\}
Step 5. Check if percentage P is more than threshold 'theta' i.e. P > 'theta'
Step 6. If Yes then allow the url 'Ur' to open
Step 7. If No then disallow the url 'Ur' from opening
Step 8. Stop

VI. RESULTS AND DISCUSSION

In our experiment we have used Google Chrome as a browser, and JavaScript, CSS, HTML to build chrome browser extension. The proposed algorithm of extension works on browser to block non-learning sites. In non-learning sites we have collected sports, entertainment, ecommerce, news, and adult sites. In learning sites we have included university, college, institutes, learning portals, educational organizations, and sites which describe different subjects like computer, science, mathematics etc.

The result analysis is carried out in two ways, i.e. using NLP and without using NLP method. We have calculated the percentage of keyword matching of each learning and non-learning sites by both methods mentioned above. As per that we have calculated true positive, true negative, false positive and false negative of proposed method. From that result calculated the precision, recall, and F1-score of the NLP-based browser extension (proposed method) and without NLP based browser extension (existing method). We have calculated all these parameters using Matlab.

Table 1 parameters of existing and proposed method

<table>
<thead>
<tr>
<th>Parameter/Method</th>
<th>TP</th>
<th>TN</th>
<th>FP</th>
<th>FN</th>
<th>FPR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing</td>
<td>565</td>
<td>563</td>
<td>239</td>
<td>233</td>
<td>0.298</td>
</tr>
<tr>
<td>Proposed</td>
<td>798</td>
<td>800</td>
<td>2</td>
<td>0</td>
<td>0.0024938</td>
</tr>
</tbody>
</table>

Table 2 Precision, Recall, F1-Score and Accuracy of both method

<table>
<thead>
<tr>
<th>Parameter/Method</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existing</td>
<td>0.70274</td>
<td>0.70802</td>
<td>0.70537</td>
<td>0.705</td>
</tr>
<tr>
<td>Proposed</td>
<td>0.9975</td>
<td>1</td>
<td>0.99875</td>
<td>0.99875</td>
</tr>
</tbody>
</table>

The graph shows the result of the proposed and existing method.

![Figure 1](https://www.ijmece.org) Result of existing and proposed method

![Figure 2](https://www.ijmece.org) Result of Accuracy

VII. CONCLUSION

Today so many parental controls are available which specially used at home or in college to protect our children from accessing harmful content while surfing on internet. To restrict students from accessing all non-learning sites is one of the challenges in this internet era. In this proposed method we got highest accuracy to block all non-learning sites. In this extension we have used natural language processing to identify the learning sites. This extension helps students to focus on their study and it also encourages students to use internet in safe and secure manner.

REFERENCES

[8] https://www.scurly.com