JCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Keyword Based Web Filtering Tool For E-Learning Sites

Sangita. S. Modi[1*], Sudhir B. Jagtap[2]

[1,2]Research Centre in Computational Science, Swami Vivekanand Mahavidyalaya, Udgir, Dt:- Latur, S.R.T.M. University, Nanded, Maharashtra, India

[*]Corresponding Author:sangitasable1@gmail.com,

Available online at: www.ijcseonline.org

Accepted: 12/Aug/2018, Published: 31/Aug/2018

***Abstract*—** The internet overwhelms us with huge amount of widely extended, well integrated, rich and dynamic hypertext information. It has deeply influenced our lives and daily routine. Billions of websites contains learning related and unrelated contents. It is very difficult to find and maintain the unrelated urls dataset to stop student from accessing the irrelevant sites in browser. Web content filtering is one of the essential tool which helps to filter out unwanted content. The proposed algorithm used to create strong keyword database of learning sites. This database used along with browser extension to analyze every incoming site and then allows browser to display only learning sites. In this extension natural language processing (NLP) plays an important role to find out and block non learning sites. We have measured the accuracy of the tool using precision and recall.

***Keywords*—**Internet, Techno-Savvy, WWW, Web Mining, Filter, NLP.

I. Introduction

In the age of internet, all type of information is available at the tip of your finger. Internet is the one of the leading technology which has become the fundamental need of every field. The education system of every county has been converted into modern education system which is pedagogical. Pedagogical means that it is a student centric education system. This system gives the importance to the feelings, opinions, interest, and learning capacity of student. This promotes student to learn anywhere, anytime, with his/her own grasping capacity. Because of internet it is possible for students to take distance education by staying at home. Information and Communication Technology (ICT) [1] is source of universal access to education, fairness in education, the availability of quality learning and teaching. ICT helps to build modern education system. It helps student to understand every concept of subject practically.

In India ICT is also introduced in all universities, colleges and schools to enhance quality of education. In that internet is one of the ICT tools which have many advantages and disadvantages. Through Internet every day students are accessing number of unwanted web sites knowingly or unknowingly. To avoid this number of filter tools are available in the market to block particular category of urls. Web filters are working on data mining [2] application called as web mining [3].

In our proposed work we have created keyword knowledge base of learning sites. Preprocessing has been done on dynamically collected web page text data using web content mining to form learning site keyword knowledge base (KKB). We have linked this KKB to the proposed browser extension to block non learning sites using natural language processing.

Our approach on web content mining, filtering and blocking will provide full-fledged dataset which will cover as much as possible keywords of unwanted sites of all categories. If this tool is to be made for college campus then, keywords based filtering can block unwanted sites. This will restrict students from taking undue advantage of high speed network to download videos, movies and also accessing social networking sites. When this tool is used, then there will be no need of urls blocking software.

In this paper section I contains introduction of proposed algorithm of keyword knowledge base creation, section II contains related work in which literature review carried out. In Section III we described methodology and step by step process of proposed system. While in Section IV, result analysis of the proposed browser extension is discussed. In the last section V we conclude the proposed system with its significance.

II. 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 [4]. Pappu et al has stated that the rapid growth of online leads to considerable outburst of data in cyberspace. The most used internet services are loaded with anonymous source and content. There is no means to supervise all web pages content on World Wide Web and therefore several web pages have poor quality [5]. 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 [6]. According to the centexitguy, 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 [7]. 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 [8]. 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 [9].

Client side filtering tool is essential tool for today's internet user. Browser based extension is one of the feasible tool which can easily installed on client. This filter requires knowledge base to filter out unwanted sites.

III. Methodology

This proposed research is constructive research in which developing a web filtration tool in the form of browser integrated extension. The proposed filter is completely keyword based.

These keywords are collected by preprocessing method. The primary data set of urls is manually collected in two categories. One is ALLOWED_URLS (AU) and another set for BLOCKED_URLS (BU). The AU set are all learning related sites which are used by student for studying different subjects, also it contains all Indian as well as foreign universities, organization sites.

AU set of learning sites are used to create keyword database. Each web page is processed to extract html content. The only text content which is going to display on site are extracted and tags are removed by regular expression. Also all numbers special symbols are also removed and only alphabetic keywords are extracted. This text data converted to tokens. After removing duplicate tokens unique learning sites keywords are collected called D'. Same process is repeated for extracting text data from BU set of non learning sites. After this the common keywords are found by taking intersection of database. The common keywords are totally removed from the D' dataset, which refine again the learning keywords database.

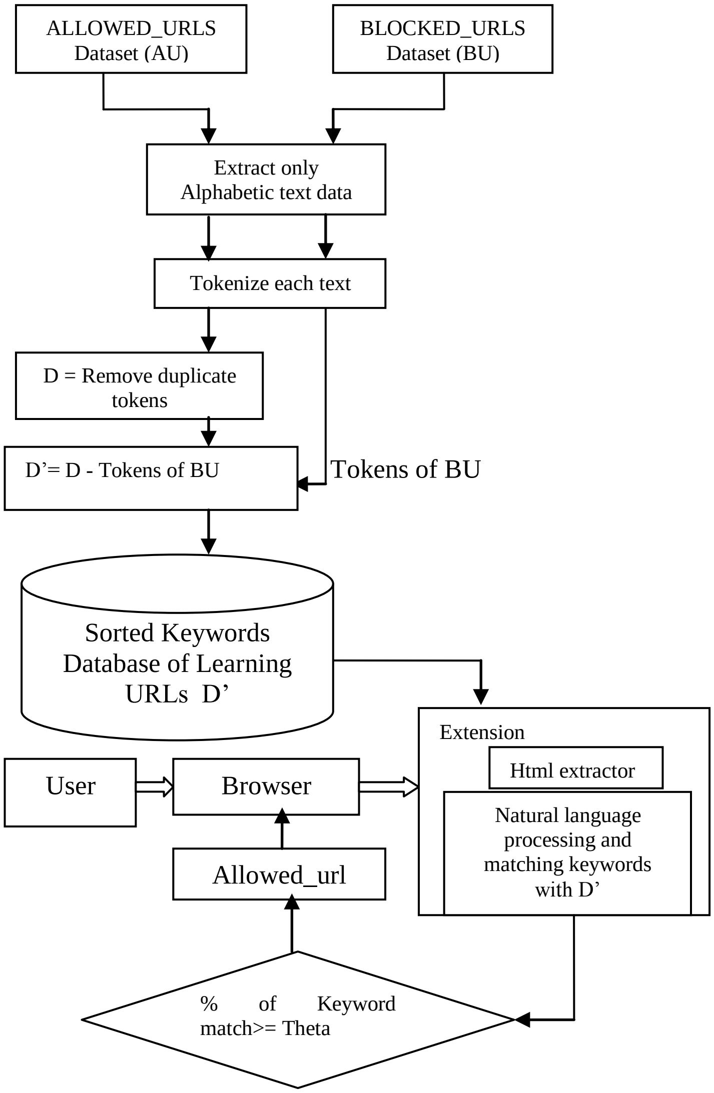
ALLOWED_URLS Dataset (AU)
BLOCKED_URLS Dataset (BU)
Extract only Alphabetic text data
Tokenize each text
D = Remove duplicate tokens
D'= D - Tokens of BU
Tokens of BU
Sorted Keywords Database of Learning URLs D'
Extension
Html extractor
Natural language processing and matching keywords with D'
User
Browser
Allowed_url
% of Keyword match>= Theta

Figure 1. Keyword Knowledgebase Used in Extension

This keyword database is used in extension to check site is learning or non learning. It uses Natural language processing to analyze current urls. The text data is extracted from the html page find the percentages of the keyword matched. If the percentage of matched keyword is greater than thresholds value then site is allowed to display in browser otherwise blocked.

IV. RESULTS AND DISCUSSION

The Matlab 2013 is used to create keyword knowledge based of learning websites. We have dynamically extracted the 2000 learning and non learning web site text contents and developed chrome browser extension using HTML, JavaScript, and CSS. The collected results are analyzed as the true positive (TP), true negative (TN), false positive (FP) and false negative (FN). The TN and TP is the outcome where the extension correctly recognizes the learning sites and non learning sites respectively. Like that FP and FN is the outcome where the extension could not recognize the learning site and non learning site respectively.

$$FPR = \frac{FP}{FP + TN} \quad (1)$$

$$TPR = \frac{TP}{TP + FN} \quad (2)$$

The two system are compared one is without NLP tool (Existing system) and another is with NLP tool (proposed system). The analyzed TPR (Recall) value is 1 and FPR value 0.0024938 as shown above equation (1) and (2).

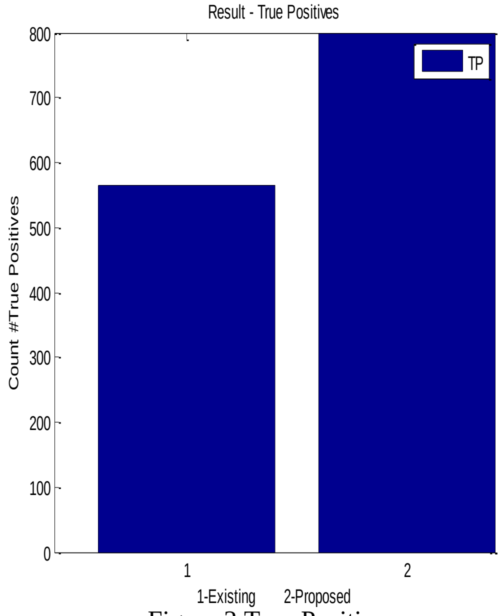

Figure 2 True Positive

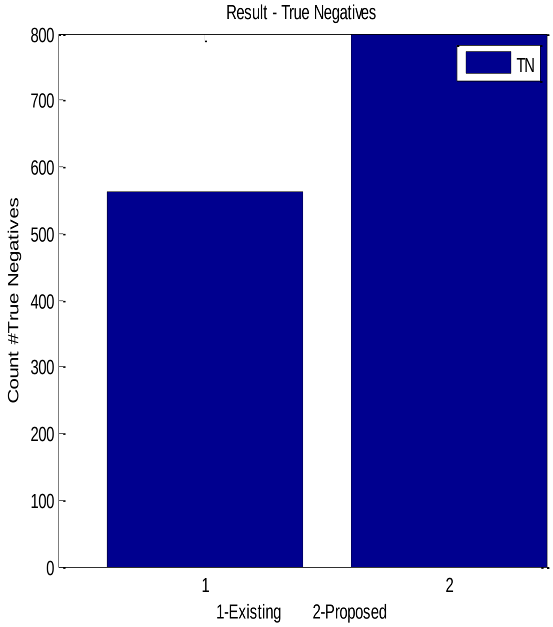

Figure 3 True Negative

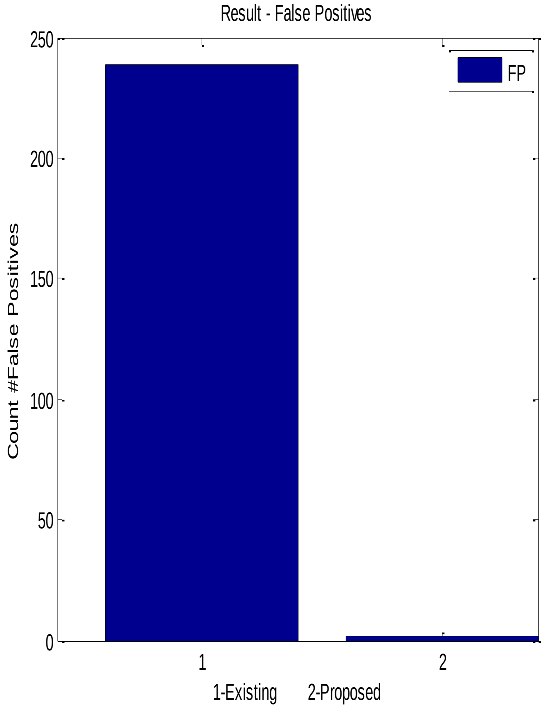

Figure 4 False Positive

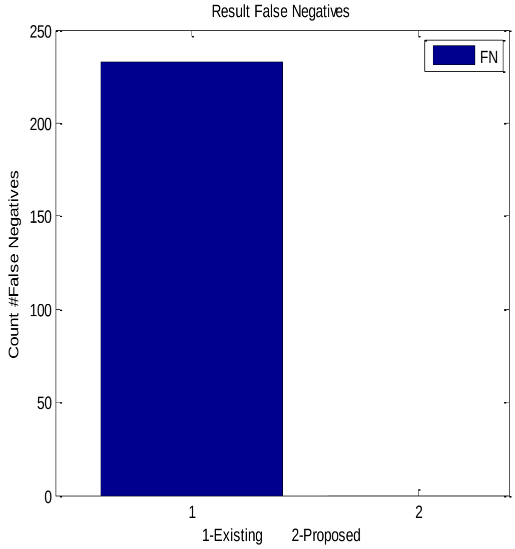

Figure 5 False Negative

V. Conclusion

The main approaches behind this browser extension development model, is keyword matching which searches the sites of strings maintained in the dataset. The main purpose of this work is to parse the document and fetch relevant keywords in the website and match these keywords with learning keyword Knowledge base. In this work we provide automatic online page assortment mechanism using chrome browser extension. The significance of the study is that it helps student to focus on their study while using internet.

References

[1] https://en.wikipedia.org/wiki/Information_and_communications_technology

[2] P. Rutravigneshwaran, "A Study of Intrusion Detection System using Efficient Data Mining Techniques", International Journal of Scientific Research in Network Security and Communication, Vol.5, Issue.6, pp.5-8, 2017

[3] A.K. Kashyap, I. Naseem , D. Mandloi (2017), "Web Mining an Approach to Evaluate the Web" International Journal of Scientific Research in Review Paper . Computer Science and Engineering Vol.5, Issue.3, pp.79-85, June (2017)

[4] Cohen-Almagor R (2015), Confronting the Internet's Dark Side, Cambridge University Press, Cambridge, pp 41R. Solanki, "Principle of Data Mining", McGraw-Hill Publication, India, pp. 386-398, 1998.

[5] Pappu A K, Trivedi A K, Sanyal S and Abraham A (2006), "SpamWall: HeuristicFilter for Web-Spam", under review in the Web Intelligence and Agent Systems Journal(WIAS), pp: 1-6

[6] Daugherty M (2004), Monitoring and Managing Microsoft Exchange Server 2003, Elsevier Digital Press, USA, pp 464

[7] Thecentexitguy (2016), Web Content Filtering: Types and Benefits, Available at http://thecentexitguy.com/web-content-filtering-types-and-benefits/, accessed on 13th February 2017.

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

[9] Reimer H, Pohlmann N and Schneider W (2015),ISSE "Highlights of the Information Security Solutions" Conference, Springer, Germany, pp 47.

Author profile

Prof. Dr. Sudhir Jagtap has completed his M.Sc., M.Phil. and Ph.D. in Computer Science from Swami Ramanand Teerth Marathwada University, Nanded. He is professor and principal of of Swami Vivekanand Shikshan Prasarak Mandal, udgir. He has more than 20 years of experience in teaching, research and administration. He has published several research papers in journals, and he is a recognized research guide in Computer Science subject of S.R.T.M. University, Nanded.

Sangita. S. Modi pursed Bachelor of computer Science and Master of computer Science from Dr.Babasaheb Ambedkar Marathwada University, Aurangabad in 1998 and 2000 respectively. She also received Master of Philosophy in computer science from yeshwantrao chavan open University in year 2012. She is currently pursuing Ph.D. in Research Centre in Computational Science, Swami Vivekanand Mahavidyalaya, Udgir, Dt:- Latur, S.R.T.M. University, Nanded, Maharashtra, India