A Keyword Based Educational and Non-Educational Website Recognition Tool

Sangita Modi, Sudhir B. Jagtap

Abstract: ***Today we all depend upon internet to do our daily activities. For booking hotel, air tickets, finding particular places, travelling, cooking, education, banking, etc. we require internet. To get a specific thing immediately, we require filtering tools. E-learning is a new and rapidly growing media in modern education system, which is totally based upon internet. While surfing on internet students may get distracted from offensive and irrelevant websites. In avoiding such distractions, filters play a vital role. This paper proposes a filter tool which carries out web scraping of text data, data cleaning, Natural language processing and filtering the non-learning sites in real-time. We have collected the text from paragraphs, images and video tags. This extracted textual data is in the form of sentences, which are processed part of speech (POS) by NLP. In NLP we are using WSD method to find the exact meaning of the ambiguous words in that context. This tool creates a knowledge base of student related sites using NLP and SVM classification technique. Word sense disambiguation is used to find the correct senses of those words, in the present sentence, which may have multiple meanings. We have created a keyword database of all learning sites. Lastly, we are classifying the sites in two categories learning and non-learning using Support Vector Machine in this tool.***

Keywords: E-learning, NLP, web content mining, SVM, POS, WSD.

I. INTRODUCTION

In today's internet era, it is very difficult to stop students from accessing unwanted data. Students always access academic or non-academic information from internet. E-learning is one of the novel approach introduced using internet. This is one of the modern education systems which has many pros and cons. Advantage is that anyone can enroll in these online courses from absolutely anywhere. But the major disadvantage is that, students may get distracted from their studies very easily while surfing on the internet. The search engine "Google" always displays a list of related and unrelated websites in its search result. Students spend a lot of time on internet to search relevant data from educational sites. It is very difficult for them to classify the listed web sites are as learning or non learning.

Today, website classification is one of the challenging tasks because; so many types of websites are coming into center of attention every year. Accessing unwanted web site is also one of the major issues arising in this era. The parental control tools are helpful for guardians and teachers to restrict students from accessing unwanted and offensive type of web sites. Browsers like Chrome, Firefox, Opera, etc. are using web site blocker extensions and plug-in and some desktop applications to block unwanted websites. Here to block offensive websites you have to give the web site's names and their keywords manually. But if any students are trying to access the website which is not present in the list, then such tools fail to block the unknown website.

Web content mining [1] is the web mining [2] technique through which we propose a filter tool which can be used to recognize the sites as learning or non learning.

A. Web content mining

Web Content Mining (WCM) is sub type of web mining used to extract the web content. Web page contains images, videos, text, banners, ads and many more things. Extraction of all type of patterns is very difficult and challenging process. In the web page source there are number of tags of HTML language. From that only text data extraction is one of the difficult tasks. Some of the web pages are not allowing such type of extraction.

After the source code reading data cleaning is essential thing through which we can get only text data by removing all tags and spaces using regular expression. Web Content Mining uses primary data of the web page. The web page content is always in unstructured form. WCM is used to identify or retrieve useful information from the structured and unstructured data. The structured data like tables, unstructured data like text and semi structured data like html document is used in web content mining. Web Content Mining is one of the difficult task, when we are processing simultaneously on the above three types of data. Web data extractor is used to extract the web content.

B. Natural Language processing

The Natural Language Processing (NLP) [15] plays a vital role in text data processing. The unstructured data like textual data is converted into structured form using features like Word occurrence, Stop Words, Latent Semantic Indexing (LSI), Stemming, N-Grams, Part of Speech, Positional Collocations, and Word Sense Disambiguation. NLP parses relatively well-formed text and sentences in different languages. Each word has at least 11 senses as nouns, 42 senses as verb. The Part-Of-Speech (POS) is used to assign the tag to each word in supervised or unsupervised manner. The WordNet is an English dictionary and associated lexical network which is used in POS tagging process.

Word Sense Disambiguation (WSD) [15] is initiated after POS tagging is completed, which is used to resolve the ambiguity of the words. WSD method is used to find the correct sense of the word, in which context it is present.

In this paper we proposed a framework which is used to identify the learning sites. The WCM is used to extract all the textual data from the HTML document.

Revised Manuscript Received on July 20, 2019.

Sangita Modi, Swami Ramanand Teerth Marathwada University, Nanded, India.

Sudhir B. Jagtap, Swami Ramanand Teerth Marathwada University, Nanded, India

Multiple pattern matching algorithms which are used to retrieve learning sites from set of learning and non learning sites. Here, we considered structured and unstructured data of the web pages. The patterns like images, video, text etc. are extracted from learning sites to form knowledge base. This knowledge base is utilized to recognize learning sites from all types of web sites. For that purpose image tag with captions, video tag with its attribute values like src, video description and alt, and text data are considered for analysis.

In this proposed system, website text content is extracted with the help of Python Beautiful Soup library. NLTK toolkit of the python is used part of speech (POS) to tag the nouns, verbs, and adjectives present in each sentence of extracted text content. The pywsd is python library which is the synonyms dataset used to find the exact meaning of those nouns which have multiple meanings called as Word Sense Disambiguation. After, all the processing we have created the dataset of learning sites keywords. This database is used to classify the sites in learning as well as non learning using support vector machine.

In this paper, Section 1 is the introduction of proposed algorithm of python based tool, Section 2 contains introduction e-learning concept. Section 3 is related work in which literature review of all web blocker and method of website blocking is described in brief. In Section 4 described the method and step by step process of our proposed tool. While in Section 5 is result analysis of the classification tool is discussed. In the last section 6 we conclude the proposed system with its significance

C. E-Learning

E-Learning is the cognitive approach of learning method by multimedia electronic learning technology. It is totally a technology dependent system, which requires infrastructure like computer, high bandwidth internet etc.. It also helps to reduce the cost of learning because, once developed a course, we may run it as many times on various locations for students. It reduces the learning time of the learners as well as they get expert knowledge, notes, and suggestions easily. If learners do not understand any concept then he/she may replay that video or re-read those notes again and again. If course is designed in an interactive way, it will help solving queries of learners.

While learning, students not only take particular online courses but also use blogs, forums, and university to gather the relevant information. Such types of sites should be considered as learning sites used by students in educational organisations as well as at home.

The disadvantage of this e-learning process is isolation and mis use of the flexibility to access internet by the learner. Learners should have high self discipline about accessing unwanted blacklisted websites in absence of parents or instructors. Website blocker is the essential filter tool which may block blacklisted websites.

D. Web Filters

A web filter is specifically used to control the website traffic. Filters are essential in blocking the web content. They are available in hardware or software form. Software filters are 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 to secure the digital assets stored in computer. There are different types of filter as follows.

1) Server Side Filter: This filter installed on server and all the clients are connected to that server. The server is responsible to monitor the network traffic.

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

3) Search Engine Filters: Search engine filter contains web crawlers which block improper websites from displaying in 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. These types of Filters work with firewall.

5) E-mail filters: It filters email headers like sender mail id, subject, and file attachments etc. to accept or reject the messages. It works in transport layer as a proxy or in application layer as a web proxy. The filtering can be customized as per user or group user requirement.

6) Social Networking filter: It is used to filter offensive posts from social networking sites. This filter mostly 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 are available 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 [3].

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 [4].

3) OpenDNS is a preconfigured Family Shield to block adult content. It works on router level [5].

4) DansGuardian blocks all images, filters ads, and blocks files from being downloaded by extension types [6].

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 [7].

6) Squid Guard is a standalone filtering tool which works at proxy level [8].

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 [9].

These available filters are used to block adult, prone, offensive, violent and irritating content of web pages. But there are no filters are available student which may open mostly learning sites, which are allowed in educational organization. This proposed filter tool specially used to recognize and create a knowledge base of the learning and non learning sites.

II. RELATED WORK

Cohen Almagor proposed 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 for their child. 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 [10]. Daugherty has stated that the client side filters have major limitation. They don't prohibit junk email before to open on user's PC. Before to use this filter the user should assure that the filter is enabled and configured [11]. According to the centexitguy the filter software needs to install on PC for content filtering called as client side filtering. The admin should decide the blocked websites list. This filters is a used for small businesses that have to control their employees [12]. The paper of Kuppusamy and Aghila has proposed work a client side filter. This filter can block the whole page or website content. It provides 88 percent accuracy in blocking the unwanted sites as well as content of that web site. The model is working as segment filter in which images and text contents are analysed[13]. Reimer et proposed that, organizations can lose control of their email easily or have to maintain and roll out solutions of client side content filtering [14].

III. FRAMEWORK OF PROPOSED WORK

This proposed Algorithm 1 is used for real-time keyword extraction from the set of Urls listed as Allowed urls and blocked urls. The allowed urls are those which are mostly used by students of any faculty for their education purpose and the blocked urls are those which contain unwanted, irrelevant data. The Algorithms are coded in python. In python for the web content mining urllib library is used. Web content like style, script and comments are omitted from extraction. After that from remaining Html source code all text displaying tags, image tag and video tag text data is extracted. If the text is in the form of sentences then that are extracted as it is.

After the extraction, all sentences are processed by Natural Language Processing. In python NLTK toolkit, the Part Of speech (POS) method is used to tag each part of sentences as nouns, preposition, verb and adjective. The Pywsd is synset dictionary of python contains the synonyms of individual nouns present in the sentences. If the any word having multiple meanings then that word is called ambiguous word.

Dataset Algorithm

The word sense disambiguation (WSD) is the process by Which we can identify the sense of word used in the sentences. A multisensory word may have homonymous or polysemous. The word like "python" having two meanings one is animal and other is programming language of computer. Suppose there are two sentences "python is good programming language in the world." and "Python is very dangerous Animal in the World". In sentence1 word "python" comes with word "programming" so it is concern with computer language, while in sentence2 it comes with animal so its meaning is as animal python.

After the processing of WSD all duplicate keywords and stop words are removed from all keyword dataset. This same process is applied on text contained along with alt attribute of image tag and on description of video tag. The output of this process is that we get three keyword dataset textual tag related keywords, image tag related keywords and video tag related keywords. Algorithm1 results two types of keywords dataset for allowed urls and blocked urls.

Step 01. Get the list of URLs URLS_Allowed = {Ua1,Ua2,...Uan}
Step 02. Get the list of URLs URLS_Blocked = {Ub1,Ub2,...Ubn}
Step 03. For each U in URLS_Allowed
Step 04. Extract text TEXT_Allowed from U
Step 05. Filter only alphabetic text i.e. TEXT_Allowed' = Fa(TEXT_Allowed)
Step 06. Tokenize TEXT_Allowed' i.e. TOKENS_Allowed = {T1, T2, ...Tm}
Step 07. Append TOKENS_Allowed to the keywords dataset DV' = {DV1,DV2,...,Dlv} for Video tags, DI' = {DI1,DI2,...,Dli} for image tags, DA' = {DA1,DA2,...,Dla} for remaining tags
Step 08. Remove duplicates DV'' = unique(DV'), DA'' = unique(DA'), DI'' = unique(DI')
Step 09. For each U in URLS_Blocked
Step 10. Extract text TEXT_Blocked from U
Step 11. Filter only alphabetic text i.e. TEXT_Blocked' = Fa(TEXT_Blocked)
Step 12. Tokenize TEXT_Blocked' i.e. TOKENS_Blocked = {T1, T2, ...Tm}
Step 13. Remove if present; TOKENS_Blocked from the keywords dataset DV'', DI'', DA''
Step 14. Sort the nouns i.e. DV = sort(DV''), DI = sort(DI''), DA = sort(DA'')
Step 15. Store the keywords dataset DV, DI, DA
Step 16. Get the list of Training Dataset URLs URLS_Training = {Ut1,Ut2,...Utn}
Step 17. Get each training url 'Ut'
Step 18. Extract text TV from the retrieved page with url 'Ut' and tag Video
Step 19. Apply NLTK to extract nouns NV = {NV1, NV2,...,NVnv}
Step 20. Find percentage of these nouns PV(i) = nv(NV Intersection DV)/nv(NV) that are present in the dataset DV = {DV1,DV2,...,DVnv}
Step 21. Check if percentage PV is more than threshold 'thetaV' i.e. PV > 'thetaV'
Step 22. If Yes then allow the url 'Ut' to open
Step 23. If No then disallow the url 'Ut' from opening
Step 24. Extract text TI from the retrieved page with url 'Ut' and tag Image
Step 25. Apply NLTK to extract nouns NI = {NI1, NI2,...,NIni}
Step 26. Find percentage of these nouns PI(i) = ni(NI Intersection DI)/ni(NI) that are present in the dataset DI = {DI1,DI2,...,DIni}
Step 27. Check if percentage PI is more than threshold 'thetaI' i.e. PI > 'thetaI'
Step 28. If Yes then allow the url 'Ut' to open
Step 29. If No then disallow the url 'Ut' from opening
Step 30. Extract text TA from the retrieved page with url 'Ut' and remaining tags
Step 31. Apply NLTK to extract nouns NA = {NA1, NA2,...,NAna}
Step 32. Find percentage of these nouns PA(i) = na(NA Intersection DA)/na(NA) that are present in the dataset DA = {DA1,DA2,...,DAna}
Step 33. Get next training url
Step 34. Pass the matrix PV, PI, PA to SVM Classifier and get the class C
Step 35. If C is Yes then allow the url 'Ut' to open
Step 36. If C is No then disallow the url 'Ut' from opening
Step 37. Stop

Filtering Algorithm

Step 01. Get the requested url 'Ur'
Step 02. Extract text TV from the retrieved page with url 'Ur' and tag Video
Step 03. Apply NLTK to extract nouns NV = {NV1, NV2,...,NVnv}
Step 04. Find percentage of these nouns PV = nv(NV Intersection DV)/nv(NV) that are present in the dataset DV = {DV1,DV2,...,DVnv}
Step 05. Check if percentage PV is more than threshold 'thetaV' i.e. PV > 'thetaV'
Step 06. If Yes then allow the url 'Ur' to open
Step 07. If No then disallow the url 'Ur' from opening
Step 08. Extract text TI from the retrieved page with url 'Ur' and tag Image
Step 09. Apply NLTK to extract nouns NI = {NI1, NI2,...,NIni}
Step 10. Find percentage of these nouns PI = ni(NI Intersection DI)/ni(NI) that are present in the dataset DI = {DI1,DI2,...,DIni}
Step 11. Check if percentage PI is more than threshold 'thetaI' i.e. PI > 'thetaI'
Step 12. If Yes then allow the url 'Ur' to open
Step 13. If No then disallow the url 'Ur' from opening
Step 14. Extract text TA from the retrieved page with url 'Ur' and remaining tags
Step 15. Apply NLTK to extract nouns NA = {NA1, NA2,...,NAna}
Step 16. Find percentage of these nouns PA = na(NA Intersection DA)/na(NA) that are present in the dataset DA = {DA1,DA2,...,DAna}
Step 17. Pass the array PV, PI, PA to SVM Classifier and get the class C
Step 18. If C is Yes then allow the url 'Ur' to open
Step 19. If C is No then disallow the url 'Ur' from opening
Step 20. Stop

This keyword dataset is used in Algorithm 2 in which we are used testing new urls set and find the given CURL is learning or non-learning. This new CURL is taken from testing urls set. This Algorithm2 extract the web content and process them as in Algorithm1. After processing all text we get three keywords set from text displaying tags, image tag and video tags. Mapping of these three dataset is carried out along with allowed keyword dataset and blocked keyword dataset of Algorithm 1. Count the percentage of that mapping. If mapping with blocked is less than mapping with allowed then as per percentage the Support vector machine classify them as Learning or Non-Learning site.

IV. RESULTS AND DISCUSSION

In our experiment we have used Google Chrome as a browser, and GUI python. The proposed algorithm is a python based tool which recognizes the Learning sites and allows opening in Chrome browser. 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 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 python tool.

$$\text{Recall} = \frac{TP}{TP + FN} \quad \text{-------(1)}$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad \text{-------(2)}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad \text{---------(3)}$$

$$\text{F1\_Score} = \frac{2X(\text{Precision}+\text{Recall})}{2X(\text{Precision}*\text{Recall})} \quad \text{\_\_\_\_\_(4)}$$

The TN and TP are the conclusions where the python based tool may recognizes truly the number of learning (class=Yes) and Non-Learning (class=No) sites respectively. While FP and FN is wrongly reorganization of leaning sites as non leaning and non learning sites learning sites. Recall (TPR) calculates how many actual Learning and Non Leaning sites exactly recognized by our tool. The recall is 0.76. Precision is the ratio of number of all positive identification of learning and Non Learning websites from the total predicted positive observations which are 0.57. The Accuracy of the correctly predicted learning and non-learning sites out of total all types of websites observation of proposed system is 0.65.

Table 1 parameter of proposed method

Parameter/Method	TP	TN	FP	FN
Proposed	35	33	26	11

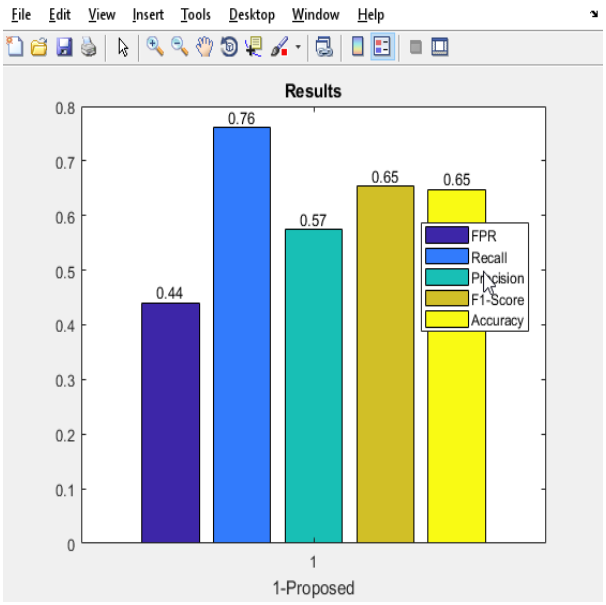

Fig. 1. Precision, Recall, F1-Score and Accuracy

V. CONCLUSION

Today so many parental controls software's are used at home, school, or in college to protect our children from accessing harmful content while surfing on internet. To restrict student from accessing all non learning sites is challenging task in today's world. In this proposed method we got highest accuracy to block all non learning sites by training the algorithm. In this tool we have used NLP part of speech to extract the nouns from the sentences and disambiguate the multiple sense nouns. This result as unique keyword dataset used to identify the learning sites. In this proposed algorithm total 1600 sites are trained. For testing 105 are used. The classifier tool's accuracy is satisfactory result. It means among different sites, maximum numbers of learning sites are recognized by this tool. This helps to create learning site knowledge base, which are useful for students.

REFERENCES

1. T. Mitchell. Machine Learning. McGraw Hill, 1997.
2. O. Etzioni. "The World Wide Web: Quagmire or gold mine", Communications of the ACM, 39(11), pp. 65–68, 1996.
3. www.k9webprotection.com
4. https://www.qustodio.com
5. https://www.opendns.com/home-internet-security/
6. http://dansguardian.org
7. http://kindergate-parental-control.com/features/block-
8. dangerous-sites
9. http://www.squidguard.org
10. https://www.securly.com
11. Cohen-Almagor R, "Confronting the Internet's Dark Side", Cambridge University Press, Cambridge, pp 41, 2015.
12. Daugherty M, "Monitoring and Managing Microsoft Exchange Server 2003", Elsevier Digital Press, USA, pp 464, 2004.
13. The Centexitguy, "Web Content Filtering: Types and Benefits", accessed on 13th February 2017.
14. Kuppusamy K S, Aghila G, "A Personalized Web Page Content Filtering Model based on Segmentation", International Journal of Information Sciences and Techniques Vol. 2, No. 1. 2012.
15. Reimer H, Pohlmann N, Schneider, "Highlights of the Information Security Solutions", Europe 2015 Conference, Springer, Germany, pp 47.
16. Preeti Dubey, "Word Sense Disambiguation in Natural Language Processing", JK Research Journal in Mathematics and Computer Sciences, Vol. 1, No. 1, March 2018.

AUTHORS PROFILE

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

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.