ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Feasibility of Using Machine Learning to Access Control on non-Educational Links

Shahenshaha Mujawar[1], Bhausaheb More[1], Pirsahab Tandel[1], Prof .Prashant Abhonkar[2]

Student, Srimatikashibainavale Sinhgad Institute of Technology and Science, kusgaon Lonavala, Savitribai Phule University of Pune, Pune, India[1]

Prof., Srimatikashibainavale Sinhgad Institute of Technology and Science, kusgaon Lonavala, Savitribai Phule University of Pune, Pune, India [2]

ABSTRACT:A number of users who use the Internet and the number of websites available on the Internet grow rapidly. Quick Internet network and billions of sites have made World Wide Web an appealing spot for individuals to utilize the Internet in their everyday life. Educational institutes provide the Internet access to students mainly for educational purposes. More often than not, students are permitted to get to any substance on the web. Therefore, the full transfer speed is expended because of access to non-educational substance, for example, streaming non-educational videos and downloading large image files, etc. Prevention of Internet usage on non-education content is practically difficult because of different reasons. For the most part, this is implemented in the proxy server through maintaining a blacklist of URLs. Most of the time, this is a static list of URLs. We have presented a methodology to dynamically prevent non-educational Internet usage for educational institutes. We have developed a machine learning data model to predict whether a given URL is educational or non educational.

KEYWORDS: Machine Learning, Squid Proxy Server, Access Control, Internet, Non-Educational Content, Streaming, Blacklist, URLs.

I. INTRODUCTION

Various clients who utilize the Internet and the number of sites accessible on the Internet develop quickly. The rising number of clients and the substance on the Internet make dealing with the Internet in an association an intricate errand for managerial clients of the system. Dealing with the Internet is a normal issue for organizations, schools and colleges. Therefore, misuse of Internet is a common social dilemma for every society. It has been found that misuse of Internet may affect individual user performance negatively. One of the common misuses of the Internet in educational institutes is accessing non-educational content such as pornography, online gaming, shopping, etc. It is observed that the majority of the bandwidth is not consumed for the intended purpose, i.e., learning. Although there are products available to prevent these misuses, they are unable to cope with the exponential growth of the World Wide Web. Squid proxy server is one such product which caches frequent usage data and can be used to access control. Most of the time, access control is done through a blacklist of URLs, which is a static methodology. Manually updating the list of URLs is time-consuming and error-prone. There are some other products such as Squid Guard which uses external databases to update the blacklist of URLs. However, frequency of updating the databases is a problem. Therefore, in this paper, we are proposing machine learning based dynamically generated blacklisted URLs for non-educational content.

II. LITERATURE SURVEY

A. Sun, E. Lim and W. Ng, [1] presented web classification using support vectorMachine. In web classification, web pages from one or more web sites are assigned to pre-defined categories according to their content. Since web pages are more than just plain text documents, web classification methods have to consider using other context features of web pages, such as hyperlinks and HTML tags.In this paper, authors propose the use of Support Vector Machine

(SVM) classifiers to classify web pages using both their text and context feature sets. Compared with earlier Foil-Pilfs method on the same data set, our method has been shown to perform very well. Authors are shown that the use of context features especially hyperlinks can improve the classification performance significantly.

A. S. Patil and B.Y. Pawar [2] proposed automated classification of web sites usingnaive bayesian algorithm.Subject based web directories like Open Directory Project's (ODP) Directory Mozilla (DMOZ), Yahoo etc., consists of web pages classified into various categories. The proper classification has made these directories popular among the web users. The exponential growth of the web has made it difficult to manage human edited subject based web directories. The World Wide Web (WWW) lacks a comprehensive web site directory. Web site classification using machine learning techniques is therefore an emerging possibility to automatically maintain directory services for the web. Home page of a web site is a distinguished page and it acts as an entry point by providing links to the rest of the web site. The information contained in the title, meta keyword, description and in the labels of the anchor (A HREF) tags along with the other content is a very rich source of features required for classification. Compared to the other pages of the website, webmasters take more care to design the homepage and it's content to give it an aesthetic look and at the same time attempt to precisely summarize the organization to which the site belongs. This expression power of the home page of a website can be exploited to identify the nature of the organization. In this paper authors attempt to classify web sites based on the content of their home pages using the Naïve Bayesian machine learning algorithm.

J. M. Pierre, [3] presented practical issues for automated categorization of web sites.In this paper authors discuss several issues related to automated text classification of web sites. They analyze the nature of web content and metadata and requirements for text features. Authors present an approach for targeted spidering including metadata extraction and opportunistic crawling of specific semantic hyperlinks. They describe a system for automatically classifying web sites into industry categories and present performance results based on different combinations of text features and training data.

R. Entezari-Maleki, A. Rezaei, and B. Minaei-Bidgoli, [4] presented comparison ofclassification methods based on the type of attributes and sample size which presents, the efficacy of seven data classification methods; Decision Tree (DT), k-Nearest Neighbor (k-NN), Logistic Regression (LogR), Naïve Bayes (NB), C4.5, Support Vector Machine (SVM) and Linear Classifier (LC) with regard to the Area Under Curve (AUC) metric have been compared. The effects of parameters including size of the dataset, kind of the independent attributes, and the number of the discrete and continuous attributes have been investigated. Based on the results, it can be concluded that in the datasets with few numbers of records, the AUC become deviated and the comparison between classifiers may not do correctly. When the number of the records and the number of the attributes in each record are increased, the results become more stable. Four classifiers DT, k-NN, SVM and C4.5 obtain higher AUC than three classifiers LogR, NB and LC. Among these four classifiers, C4.5 provides higher AUC in the most cases.

J. MorahanMartin [6] presented internet abuse: emerging trends and lingering questions.Concern about Internet abuse has grown as Internet use has proliferated worldwide. Although some question whether IA exists, clinics have been established to treat those with IA and a growing body of research over the last decade has documented that, worldwide, a small percentage of Internet users develop disturbed patterns of behavior that have been called Internet abuse. People with mood disorders and those who are lonely, socially anxious, or use the Internet to cope with negative feelings are vulnerable to IA. Some researchers conceptualize IA as a continuum of deficient self-regulation that ranges from normal to disturbed use, while others conceptualize IA as a clinical disorder. Even in those cases in which IA reaches clinical significance, there is disagreement about whether IA is a distinct disorder. Some view IA as symptomatic of other, primary disorders that are frequently co-morbid with IA, such as mood disorders or social anxiety. Among those who contend that IA is a clinical disorder, the consensus is that IA should be considered as an impulse control disorder NOS, that is, individuals' habitual inability to control their Internet use, which causes clinical levels of distress or impairment. Problems with impulse and particularly impulse control disorders have been associated with behavioral addictions. Although behavioral addictions are not universally recognized, there is a growing body of evidence that disturbed patterns of substances and some behaviors – that is, chemical and behavioral addictions – share similar features and etiology. This model may apply to IA. Cognitive behavioral models of IA have also been proposed. Some users develop IA from problematic use that is specific to a given online application (e.g., MMORPGs) or behavior (e.g., downloading pornography). These are called specific IAs. However, many users develop problems from what has been called a generalized form of IA. These users prefer socially interactive aspects of the Internet. Many are less inhibited in their online social interactions and develop a preference for online over F2F interaction. There is overlap between some specific IAs and generalized IA. In both types, social uses of the Internet or using the Internet to cope with stress become problematic. Future research should focus on the similarities and differences between specific and generalized IA. Although many clinicians report that they have treated clients with IA or specific IA, few have training in recognizing and treating IA. Research on treatment of IA is limited. IA is a relatively new area of research, and many areas need further research. The lack of a uniform set of empirically validated criteria for IA is a weakness that pervades much of the research of IA. This is especially pertinent for those who suggest that IA should be accepted as a clinical disorder. Most studies have been surveys of adolescents and university students, which limits generalizability.

K. S. Young and C. J. Case, [7] presented employee internet management: current business practices and outcomes. This paper empirically examines emergent business practices that attempt to reduce and control employee Internet misuse and abuse. Over a 6-month period, 52 web-administered surveys were collected. Respondents ranged from human resource managers to company presidents. Data were stored in a database management system and analyzed utilizing statistical measures. Monitoring efforts and policy development issues are examined against critical incidents of employee Internet abuse. The analysis also includes a rank ordering of the types of Internet applications that were perceived as most problematic or abused. Types of applications abused include electronic mail, adult web sites, online gaming, chat rooms, and stock trading, and so on. Moreover, company size and years online are examined. Overall, this research will assist organizations in implementing effective corporate initiatives to improve employee Internet management practices.

X. Qi and B. D. Davison,[8] proposed web page classification: features and algorithms.Classification of Web page content is essential to many tasks in Web information retrieval such as maintaining Web directories and focused crawling. The uncontrolled nature of Web content presents additional challenges to Web page classification as compared to traditional text classification, but the interconnected nature of hypertext also provides features that can assist the process.As authors review work in Web page classification, they note the importance of these Web-specific features and algorithms, describe state-of-the-art practices, and track the underlying assumptions behind the use of information from neighboring pages.

III. SOFTWARE REQUIREMENT SPECIFICATION

User Classes and Characteristics

To design products that satisfy their target users, a deeper understanding is needed of their user characteristics and product properties in development related to unexpected problems that the user's faces every now and then while developing a project. The study will lead to an interaction model that provides an overview of the interaction between user characters and the classes. It discovers both positive and negative patterns in text documents as higher level features and deploys them over low-level features (terms).In proposed work is designed to implement above software requirement. To implement this design following software requirements and hardware requirementsare used.

Software Requirements

➢ Operating System	-	Windows XP/7
➢ Programming Language	-	Java/J2EE
➢ Software Version	-	JDK 1.7 or above
➢ Tools	-	Eclipse
➢ Front End	-	JSP
➢ Database	-	Mysql

Hardware Requirements

➢ Processor	-	Pentium IV/Intel I3 core
➢ Speed	-	1.1 GHz
➢ RAM	-	512 MB (min)
➢ Hard Disk	-	20GB
➢ Keyboard	-	Standard Keyboard
➢ Mouse	-	Two or Three Button Mouse
➢ Monitor	-	LED Monitor

IV. COMPARISON BETWEEN EXISTING SYSTEM AND PROPOSED SYSTEM

Existing system uses Squid proxy server. It is one such product which caches frequent usage data and can be used to access control. Most of the time, access control is done through a blacklist of URLs, which is a static methodology. Manually updating the list of URLs is time-consuming and error-prone. There are some other products such as Squid Guard which uses external databases to update the blacklist of URLs. However, frequency of updating the databases is a problem.

We propose feasibility of using machine learning to access control on non-educational links. We propose a methodology to generate dynamic blacklist of URLs using machine learning techniques. We have developed a machine learning data model to predict whether a given URL is educational or non educational. The methodology described in this paper can be used to block the non-educational content dynamically. Our system should not be overhead for the existing system. Therefore, the latency in our solution should be minimized. The proposed system will not block any new URLs. The system takes a copy of the request from the server log with the response and then it processes in the background. If the system identifies the request as non educational, then the system will update the blacklist. When a user tries to access the same URL, access will be denied by the system. The frequency of the update could be decided by the administrative user of the system.

V. ALGORITHM FOR RELEVANT FEATURE DISCOVERY

- **Seed Collection:**

First we check user request is URL or search query. If it is search query then we remove stop words and performing seed collection. Seed collection means we get words related URLs online.

- **Reverse Searching:**

After performing seed collection we heck words are presents in URLs or not. If search word is present in URL then we consider as relevant links otherwise that link is irrelevant. After classification of relevant and irrelevant we check URLs, URL title and URL description with education related word dictionary and then again classify that particular URL is relevant or irrelevant.

VI. SYSTEM ARCHITECTURE

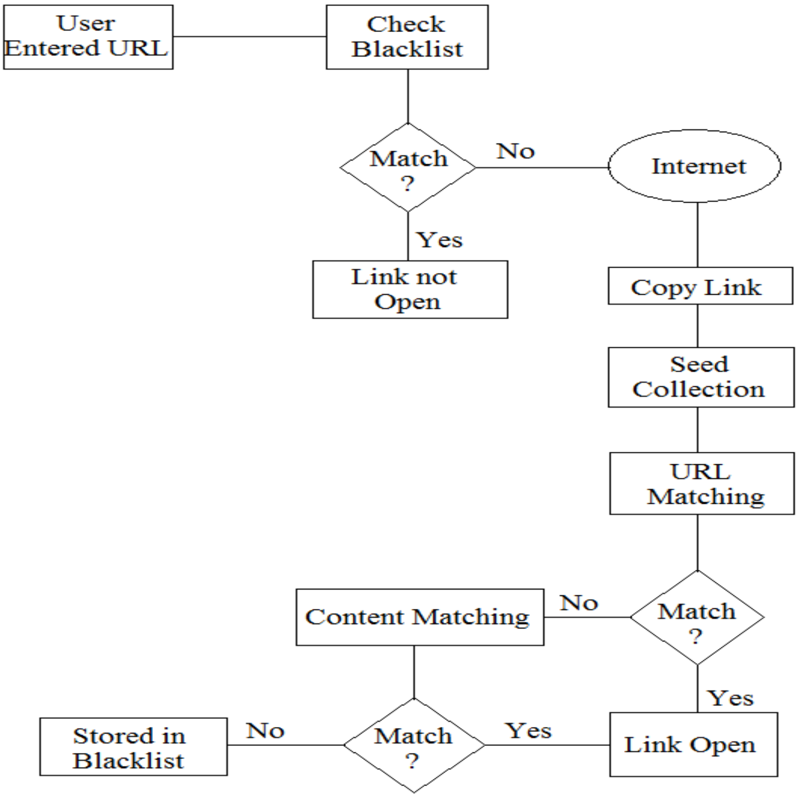


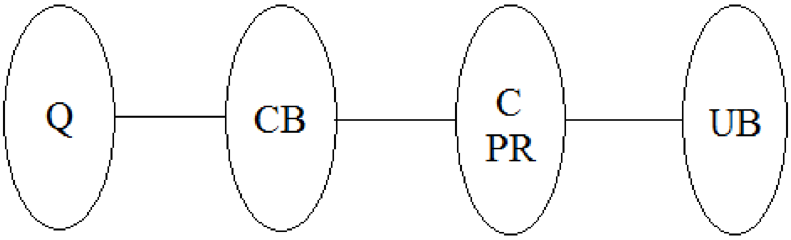
Figure 1: System Architechture

DOI: 10.15680/IJIRCCE.2018.0605050

Above diagram shows the overall architecture of the proposed system. When a user makes a request of a website, the request is checked against the blacklist. If the blacklist allows the request, it sends to a remote server, and the response will be given to the user. The Copy Request component will copy all the requests by the users. Then Preprocess Request component will extract the requested URLs. Then we performed seed collection and then perform URL matching.If URL is matched with educational contents then that URL is open and if URL is not matched then we perform content matching.Finally, if contents are matched with educational contents then only that URL is open otherwise URL is not opened and stored in URL blacklist.Finally, URL and the decision of the data model will be stored in the database.

VII. MATHEMATICAL MODULE

A] Mapping Diagram



Where,

Q = User entered query i.e. URL
CB = Check blacklist
C = Copy link i.e. URL
PR = Preprocess request
UB = Update blacklist

B] Set Theory

1) Let S be as system which find entered URL is blacklist or not.
S= {In, P, Op, Φ}
2) Identify Input In as
In= {Q}
Where,
Q = User entered query i.e. URL
3) Identify Process P as
P= {CB, C, PR}
Where,
CB = System check entered list is present in blacklist or not
C = Copy that link i.e. URL
PR = Preprocess request
4) Identify Output Op as
Op = {UB}

Where,
UB = Update blacklist

After preprocessing the request, system decides particular link is education related or not. If it is not educational related then system add that link into blacklist.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Website: www.ijircce.com

Vol. 6, Issue 5, May 2018

Φ = Failures and Success conditions.

Failures:

1. Huge database can lead to more time consumption to get the information.
2. Hardware failure.
3. Software failure.

Success:

1. Search the required information from available in Datasets.
2. User gets result very fast according to their needs.

Space Complexity:

The space complexity depends on Presentation and visualization of discovered patterns. More the storage of data more is the space complexity.

Time Complexity:

Check No. of patterns available in the datasets= n

If (n>1) then retrieving of information can be time consuming. So the time complexity of this algorithm is O(n^n).

VIII. EXPERIMENTAL SET UP AND RESULT TABLE

Result Table:

Sr. No	Search Query Id	Time in milliseconds
1	1	300
2	2	190
3	3	120
4	4	250
5	5	800
6	6	60

Table 1: Search query execution time

Result Graph:

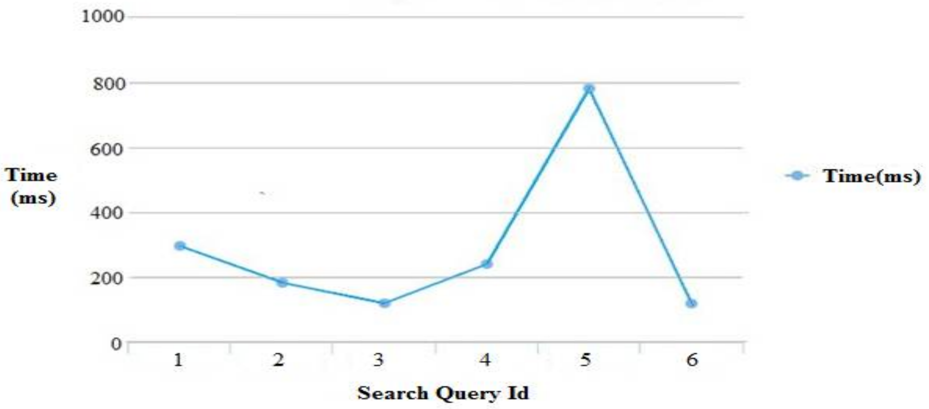


Figure 2: Search query execution time

DOI: 10.15680/IJIRCCE.2018.0605050

Above Figure 2 shows the time required for execution of each query. In above graph X-axis represents search query Id and Y-axis represents time in milliseconds. Above graph shows that our system works effectively and efficiently and also shows that our system works better as compare to other state of art systems.

IX. CONCLUSION

We have exhibited feasibility of using machine learning to access control on non-educational links. Prevention of Internet usage on non-education content is practically difficult due to various reasons. Usually, this is implemented in the proxy server through maintaining a blacklist of URLs. Most of the time, this is a static list of URLs. With the fast growing content on the World Wide Web maintaining a static blacklist is impractical. In this paper, we have presented a methodology to dynamically prevent non-educational Internet usage for educational institutes. We have developed a machine learning data model to predict whether a given URL is educational or non educational. The methodology described in this paper can be used to block the non-educational content dynamically.Our system resolve manually updating URL list problem. Our proposed system gives best performance as compare to other state of art systems.

REFERENCES

[1] Sun, E. Lim and W. Ng, "Web classification using support vector machine", in Proc. of the 4th Int. workshop on Web information and data management, McLean, Virginia, USA, 2002, pp. 96-99.

[2] S. Patil and B.Y. Pawar, "Automated Classification of Web Sites using Naive Bayesian Algorithm", in Proc. of the 4th Int. workshop on Web information and data management, March 14 - 16, 2012, Hong Kong.

[3] J. M. Pierre, "Practical issues for automated categorization of web sites", in Electronic Proc. of ECDL 2000 workshop on the Semantic Web, Lisbon, Portugal, 2000.

[4] R. Entezari-Maleki, A. Rezaei, and B. Minaei-Bidgoli, "Comparison of Classification Methods Based on the Type of Attributes and Sample Size", Journal of Convergence Information Technology (JCIT), Vol. 4, No.3, pp. 94-102, 2009.

[5] T. H. Witten, E. Frank and M. A. Hall. "Data Mining: Practical Machine Learning Tools and Techniques", Morgan Kaufmann, San Francisco, 2011.

[6] J. MorahanMartin. "Internet abuse: Emerging trends and lingering questions." In A. Barak (Ed.), Psychological aspects of cyberspace: Theory, research, applications (pp. 32-69). Cambridge, UK: Cambridge University Press, 2008.

[7] K. S. Young and C. J. Case, "Employee Internet Management: Current Business Practices and Outcomes", CyberPsychology and Behavior, 5(4), p 355-361, 2002.

[8] X. Qi and B. D. Davison, "Web Page Classification: Features and Algorithms", ACM Computing Surveys, Vol 4 1, No. 2, Article 12, 2009.