International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 12 | Dec 2019 www.irjet.net p-ISSN: 2395-0072

Noisy Content Detection on Web Data using Machine Learning

Dr. M.C Hingane[1], Sharad Hake[2], Jugal Badgujar[3], Ravindra Shingade[4], Rohit Jadhav[5]

[1]Assistant Professor Dept of Computer Science PDEA's COEM
[2,3,4,5]Student's Dept of Computer Science PDEA's COEM

Abstract -*The World Wide Web holds a huge amount of data. The web is doubling in size every six to ten months. The World Wide Web facilitates anyone to upload and download relevant data and the precious content on the web site can be used in all fields. The website has become the intruder's main target. An intruder embeds Noisy content in web pages for the purpose of doing some bad and unwanted activities. That noisy content includes advertisements, known relevant data that does not use for the user. Whenever user retrieve any information from the web side it also brought some noisy content. Web mining is one of the mining technology which mines the data in large amount of web data to improve the web service.*

Key words - Machine learning, web mining, Classification, Convolution Neural Network, Decision Tree

1. INTRODUCTION

Information from the web, mining techniques are used A web site is a collection of related web pages containing images, videos or other digital assets. In fact, in today's age, it is almost mandatory to have an online presence to run a successful e in any everyone to upload and download the information. That information can use in any filter. But it also had a disadvantage there technological elevation come coupled with new sophisticated technologies to attack and scam users there are also some noisy data in web data the noisy data contents are

- Advertisement
- Irrelative data

Some of the offending word

The noisy data like an advertisement it's also called a Malicious content Attacker place their advertisement on the web page. If the user clicked on that advertise then it opens another web page that holds the harmful contents. It is just like a phishing attack. In order to, for example, understand user behavior or reset of search engineer it is necessary to and use the information available on the web. These fields that describe these tasks are called web mining. The web holds very largely and periods access to it from any chance and any time. Most people browse the internet forget information, but most of the time, they get an important and irrelevant (related) document each after navigation several links. For retrieving info

2. PROBLEM IDENTIFICATION

Web usage mining is the application of data mining ways of doing things to web clickstream data in order to extract useful data. As website continues to grow in size and complex difficulty, the result of web usage mining has become critical for some application such as web site design, Basically, there are 2 challenges are involved in the web mining and that are processing the raw data used to provide a (very close to the truth or true number) pictures which shows that how the site is being used, and results of the different data mining set of computer instruction in order to present the only rule and pattern are filtered

For the mining data from web log files, an effective and efficient algorithm is required that works with high performance. Moreover, it required to authenticate the algorithm for that purpose we use a traditional algorithm for mining sequential pattern from weblog data. In this work, the author develops decision tree algorithm, which is efficient mining method to mine log files and extract knowledge from the web data stream and generated training rules and Pattern which are helpful to find out different information related to logging file. The author increases the accuracy of generating non-redundant association rules for both nominal and numerical data with less time complexity and memory space. In this method, the Author uses the N-fold cross-validation technique for performance evaluation and for classification of data set author is using a decision learning algorithm with some modification in the decision tree algorithm.

3. RELATED WORK

URL filtering compares the URLs that end-users attempt to access to URL categories or lists of URLs. You can use URL filtering to prevent users from accessing websites that provide content that is objectionable, potentially harmful, or not work-related. This kind of content filtering can increase network security and enforce an organization's policy on acceptable use of resources. In URL filtering, the engines compare the URLs in HTTP and HTTPS requests

against URL categories or lists of URLs. There are two ways to define the URLs:

- You can use URL Category and URL Category Group elements to filter URLs based on URL categorization.
- You can use URL List elements to filter specific URLs.

You can use both methods together. You can also define allowed URLs manually if a URL that you want to allow is included in a category of URLs that you otherwise want to block.

The URL categorizations are provided by the external Force point™ Threat Seeker® Intelligence Cloud service. Threat Seeker Intelligence Cloud (Threat Seeker) provides categories for malicious websites and several categories for different types of non-malicious content you might want to filter or log. The NGFW Engine sends categorization requests using HTTPS to Threat Seeker.

URL Category Group elements contain several related URL Categories. When you use URL Category Group elements in the Access rules, the rule matches if any of the URL Categories included in the URL Category Group match.

The engine can use the server name indication (SNI) in HTTPS traffic for URL categorization without decrypting the HTTPS connection. When a web browser contacts a server to request a page using HTTPS, the browser sends the server name in an unencrypted SNI field. However, the requested URL is not known when HTTPS connections are not decrypted. How

You can use category-based URL filtering to create Access rules that disallow decryption for traffic in the specified categories. For example, you can create rules that prevent traffic to online banking services from being decrypted.

You can configure the engine to respond in various ways when a match is found. For example, you can log the matches or block the traffic. If you decide to block traffic, the engine can notify end users with a message in their browsers. You can define customized User Response elements for URL filtering matches, such as a custom HTML page that is displayed to the end user when a connection is blocked. If the engine detects that it cannot connect to Threat Seeker, all URLs match the **Data Provider Error** URL Category. You can optionally add Access rules to discard all traffic that cannot be categorized

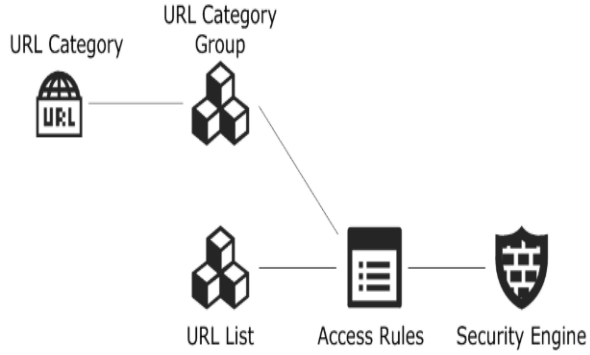


fig 3.1 element in the configuration

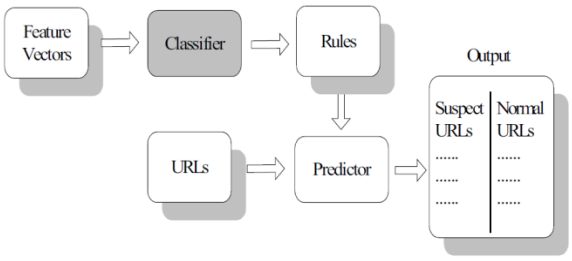


fig 3.2 Url Classification pattern

4. Proposed Methodology

The present invention overcomes the deficiencies of the prior art with a system for web-based content detection in image, videos, text and image extraction via a browser plug-in. The system is advantageous because it performs initial content processing at a client to determine whether an content is suitable for a recognition process and extracts an content frame for processing

Machine Learning (ML) is a research area within Artificial Intelligence (AI) and Statistics concerned with the automatic acquisition of knowledge from data sets, studying techniques able to improve their performance from experience [3]. One of the main areas of ML is data classification, where classifiers are induced using a training set to assign new, previously unseen examples to their correct class. The classification techniques employed in this work follows different learning paradigms, presenting distinct bias. This choice was made such that different predictors could be ensured, improving their combination for noise detection. The following ML algorithms were chosen:

- Support Vector Machines,
- Artificial Neural Networks,
- Decision Tree

- k-nearest neighbor
- Naive Bays

a) Support Vector Machines (SVMs): are based on the Statistical Learning theory. They split data from two classes with a maximal margin hyper plane, which can then be used in the classification of new data [4]. In their algorithm we want to classify the noisy content.

b) **Artificial Neural Networks (ANNs):** are composed of simple processing units, simulating the biological neurons, which are named nodes of the ANN [5]. The ANN training consists of adjusting weights of connections between the artificial nodes, and its able to predict the content are safe or not

c) **Decision Tree (DT):** induction algorithm. A DT is composed of leaf nodes, representing the classes, and by decision nodes, representing tests applied to the values data attributes [6]. The classification of new data is performed by traversing the tree according to the results of the tests until a leaf node is reached.

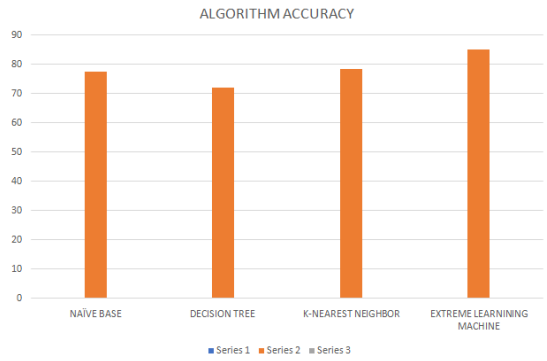
d**) k-Nearest Neighbor (KNN):** is an instance-based technique where the classification function is approximated locally in order to obtain predictions for new data in this section crawler is able to find path on base of past result.

e) Naive Bayes - Naive Bayes classifiers are a collection of classification algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e. every pair of features being classified is independent of each other

In these module we have more option to choose right algorithm. According to model classification we can also use page rank sum text for giving a rank to page

There we classify the algorithms ata form of the histogram there we see which algorithm is best for the our model but at the end k nearest neighbor are most use In out model because its give a maximum percentage result

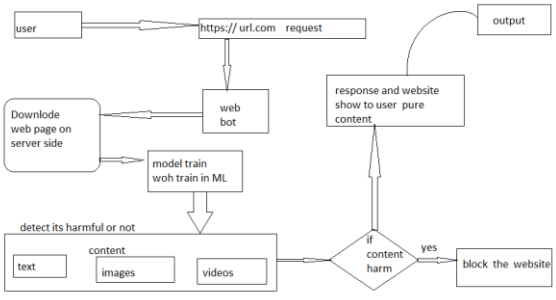
According to this give machine algorithm see the accuracy at view of histogram



5. Workflow

We built WebGuard over a client-server architecture, shown in Figure 1. Users access WebGuard over the Internet through a. The query processor communicates with a feature database and the knowledge base, where the system stores semantic and fact-based metadata, respectively. The web filter system (WebGuard) aims to block those sites. It provides Internet content filtering solutions and Internet blocking of noisy data and many more categories. The Internet will thus become more controllable and therefore safer for both adults and children.

There is a general consensus regarding certain types of web sites that they must be "filtered" and "blocked" so children do not inadvertently gain access to them. A number of different organizations have created their own definitions of what is or is not appropriate for children using the Internet. The following categories1 are intended to only to serve as a guideline based on the types of materials, text and pictures currently on the Internet



In this Architecture we want to block web site on bases of the noisy content where in this section first use want to request for the url search where on old model we block the site using black list and white list format where in that we already uploaded the url and then the its search url is this

harmful or not if that is block the site but in our model we create a web bot

Who automatically detect the noisy content so after user request web bot is download the page at background and then using some knn, svm type algorithm it detect the noisy data one more thing in this model we train the model for classify the content if that content is harmful for children or adult then we block the site in that section we also remove the ad maximum type of data we can remove

6. Conclusion & Future Work

The paper suggested an approach to the classification of texts, images data for inappropriate content blocking in the Web using Data Mining techniques. Architecture and algorithms of text classification were considered. The approach to improve the TF/IDF approach for the text classification was outlined. The proposed approach for combining simple (binominal) classifiers, which are responsible for decision whether an object belongs to one specific category or not,. We also go for the mining from cloud. Whenever we work on mining over cloud computing that time we hesitate for the cost but that come very less by cloud mining. So, we can say that cloud mining can seen as future of web mining

7. RESULT

Web is a rapid growing research area it consists of Web usage Web structure and the Web content mining. Web usage mining refers to the discovery of user access patterns from Web usage logs. Web content mining aims to extract/mine useful information or knowledge from web page contents

In our paper we try to identify the problems path and prevent the unauthorized accessibility and we also care about the user similarly in than paper we also try to focus show the user pure content not impure content and what's they need

8. REFRENCE

[1] Igor Kotenko 1, Andrey Chechulin "Evaluation of Text Classification Techniques for Inappropriate Web Content Blocking rules,".

[2] Arvind Kumar Sharma1, P.C. Gupta "Study & Analysis of Web Content Mining Tools

[3 M. Aldekhail, "Application and Significance of Web Usage Mining in the 21st Century

[4] Gowtham Mamidisetti, Nalluri Gowtham, Ramesh Makala "Web Data Mining Framework for Accidents Data".

[5] S. T. Dumais, J. Platt, D. Heckermann, M. Sahami, "Inductive learning algorithms and representations for text categorization,"

[6] K. Nethra1, J. Anitha2 and G. Thilagavathi3, "WEB CONTENT EXTRACTION USING HYBRID APPROACH

[7] T. Joachims, "Text categorization with support vector machines: learning with many relevant features," in *Proceedings of the 10th European Conference on Machine Learning*, Chemnitz, Germany, 1998, pp. 137-142.

[8] Raymond Kosala"Web Mining Research: A Surveyblocking,"

[9] I. V. Kotenko, A. A. Chechulin, A. V. Shorov, D. V. Komashinkiy, "Automatic system for categorization of

[10] Hyunsang, Heejo Lee" Detecting Malicious Web Links and Identifying Their Attack Types".

[11] DNS-BH. Malware prevention through domain blocking. http://www.malwaredomains.com.

[12] FETTE, I., SADEH, N., AND TOMASIC, A. Learning to detect phishing emails. In WWW: Proceedings of the international conference on World Wide Web (2007).

[13] GARERA, S., PROVOS, N., CHEW, M., AND RUBIN, A. D. A framework for detection and measurement of phishing attacks. In WORM: Proceedings of the Workshop on Rapid Malcode (2007).,

[14] GEOIP API, MAXMIND. Open source APIs and database for geological information. http://www.maxmind.com

[15] pgAdmin, http://www.pgadmin.org/, last accessed on 03.03.2015.

[16] PostgreSql, http://www.postgresql.org/, last accessed on 03.03.2015.

[17] X. Qi and B.D. Davison, "Web Page Classification: Features and algorithms," *ACM Computing Surveys (CSUR)*, vol. 41, issue 2, 2009, article no. 12.

[18] RapidMiner https://rapidminer.com/. last accessed on 03.03.2015.

[19] P. Schauble, *Multimedia Information Retrieval: Content-Based Information Retrieval from Large Text and Audio Databases*, Kluwer, MA, USA, 1997.

[20] M. Tsukada, T. Washio, H. Motoda, "Automatic web-page classification by using machine learning methods," *Lecture Notes in Computer Scince*, Springer, vol. 2198, 2001, pp. 303-313.

[21]Yandex.TranslateAPI
http://api.yandex.com/translate/, last accessed on 03.03.2015.