RESEARCH ARTICLE OPEN ACCESS

A survey on Web Contents Classification System

Prof. Nisha R. Wartha[1], Prof. Vaishali Londhe[2]
[1]Professor in Information Technology Department Government Polytechnic, Thane,
[2]HOD of Computer Engineering DepartmentYadavraoTasgaonkar Institute of Engineering and Technology.

Abstract:

Currently, the amount of adult (pornographic) content on the Internet is increasing rapidly. This makes an automatic detection of adult content a more challenging task, when eliminating access to ill-suited websites. It is easy for children to access pornographic webpages due to the freely available adult content on the Internet. It creates a problem for parents wishing to protect their children from such unsuitable content.In 2005, the European Parliament launched a large program called "Safer Use of the Internet", particularly for young people. Some webpages contain a huge amount of combined data related to healthcare (information on diseases, mental health, and physical fitness) and sexual knowledge (medicine for sexual health, birth control, treatment during pregnancy, etc.). In this system, we focus on the recognition of Web adult content. A fuzzy-ontology/SVM–based adult content detection system is proposed to automate the classification of pornographic versus medical websites. The proposed mechanism offers an adult content detection system that classifies webpages into normal, pornographic, or medical webpages using extracted web content features. The adult Web page bag recognition is carried out using multi-instance learning based on the combination of classifying texts, images and videos in Web pages.

Additional Key Words and Phrases: Semantic knowledge, fuzzy ontology, SVM (Support Vector Machine), adult content identification, skin patch modeling, recognition of adult images, recognition of adult videos, recognition of adult Web page bags,k-NN (nearest neighbor).

Keywords — **Put your keywords here, keywords are separated by comma.**

I. INTRODUCTION

With the enormous growth of the World Wide Web (WWW), there are more and more websites providing information often considered offensive and obscene. One can easily find pornographic sites using global web search engines. A huge number of adult webpages on the internet are freely available to all users, which can damage the mental and physical health of teenagers [15]. It also creates problems for parents wishing to keep children away from these webpages [16]. No doubt it's necessary to protect children from adult content.

In addition, some webpages contain a huge amount of combined data related to healthcare (information on diseases, mental health, and physical fitness) and sexual knowledge (medicine for sexual health, birth control, treatment during pregnancy, etc.).

Different methods have been introduced to block or restrict access to adult websites such as IP address blocking, text filtering, and image filtering. The Internet Protocol (IP) address blocking bans the adult content from being accessed by certain users.

This technique works by maintaining a list of IPs or Domain Name Servers (DNS) addresses of such non-appropriate websites. For each request, an application agent compares the requested website IP address or DNS with the restricted list. The request is denied if the two addresses match, and approved otherwise. This method requires manual keeping and maintenance of the restricted list IPs, which is difficult as the number of the adult content websites grows or some websites change their addresses regularly.

Filtering by text is the most popular method to block access to adult content websites. The text filtering method blocks the access to a website if it contains at least one of the restricted words. Another approach is to use a machine learning algorithm to find the restricted words. Sometimes, instead of using the machine learning technique to extract keywords, a classification model is used directly to decide whether the requested webpage is safe [14]. Nonetheless, the text blocking method only understands texts, and it cannot work with images and videos. This problem arises when the webpage does not contain the restricted keywords or does not contain text at all. As well as, it may block safe webpages such as a medical webpage as it contains some restricted keywords.

II.LITERATURE REVIEW

Text classification has been used to recognize Web adult information. As following literature review shows that multiple researcher works on web page text to recognition of adult text.

[1]Farman Ali , Pervez Khan, KashifRiaz, DaehanKwak, Tamer Abuhmed, Daeyoung Park, Kyung Sup Kwak[2017] proposed a A Fuzzy Ontology and SVM–Based Web Content Classification System which classifies the provided URLs into adult URLs and medical URLs by using a blacklist of censored webpages list to provide accuracy and speed. The proposed fuzzy ontology then extracts web content to find website type (adult content, normal, and medical) and block pornographic content.

[2]Du et al. [2003] proposed a Web filtering system that uses a text classification algorithm to classify web pages into adult and non-adult pages.

Information extraction and pornographic content filtering are sensitive topics in the field of information engineering research [3]–[10]. The increase in adult websites on the Internet has made web filtering a more challenging task. Most pornographic website filtering systems are unable to filter data efficiently to prevent teenagers from accessing them. A solution to existing problem, comprehensive technological work is required to extract and filter adult contents from the web data

intelligently and deny access to ill-suited webpages systematically. One possible method of webpage filtering is to record the URLs of ill-suited websites. The main advantage is speed. However, a URL based filtering system does not work perfectly every time, since many URLs do not present the actual information. To handle this limitation, web content-based filtering and blocking techniques are required to filter webpages competently. Different links and Hypertext Markup Language (HTML) tags in webpages contain a lot of information that can be used for filtering.

In contrast to adult texts, adult images are considered to be more influential, because image information is much more rapidly perceived and the graphic effects are often more shocking and disturbing to people. There have been many images and video-based filtering and blocking techniques are also used to stop access to unsuitable webpages. Skin detection is usually usedas the most common parameter for detecting and blocking obscene images. Adult image classification methods use two kinds of filter; an adult image filter and a harmful symbol filter [11]. The adult image filter uses a statistical model for skin detection and a neural network for adult image classification. The experimental results with both filters showed promising performance. A novel framework for webpage splitting handles three categories of webpages. These are a continuous text classifier, a discrete text classifier, and a fusion-of-images classifier [12]. These classifiers provide a decision symbol (porno or non-porno) to the web browser.

Adult videos are often considered to be more influential than adult images, due to their increasedrealism and ability to portray actions in detail. There have been several attempts to build adult video detection systems. [13] Adult Image and Video Recognition by a Deep Multicontext Networkand Fine-to-Coarse Strategy proposed a deep multicontext network with fine-to-coarse strategy for adult image and video recognition. They employ a deep convolution networks to model fusion features of sensitive objects in images. For diverse target objects, a novel hierarchical method is investigate and a task-specific strategy is designed, which make the multicontext method more suitable for adult Images and Videos recognition.[15] This paper proposes ACORDE, a novel deep learning architecture that comprises both convolutional neural networks and LSTM recurrent networks for adult content detection in videos.

.

III. PROPOSED SYSTEM

A. Existing System-

Many adult content webpage–filtering systems available which are based on different methods such as n-gram, naïve Bayes andkeyword-matching mechanisms.These systems have no

reasoning capability to intelligently filter web content to classify medicalwebpages from adult content webpages. Current systems for content filtering are Keyword-based searching which is a renowned form of search engine query to easily retrieve the data of these webpages. It retrieves the data by comparing the query keywords with web content words and returns the result. However, the existing adult content filtering systems are inefficient at detecting whether the webpage is about pornography or medicine.

At present, the detection of adult content is based on name of website that is uniform resource locator (URL) filtering, image filtering, and some dynamic filtering mechanisms exists. URL filtering methods use URL blacklists and do not evaluate the content of a webpage, which increases the possibility of a wrong decision. Image filtering techniques might identify medical-related images as adult images, and they don't have high accuracy.

A dynamic filtering system analyzes the content of a webpage using various algorithms. Like Naive bayes or classical ontology, which has thousands of keywords to compare and classify the content. Although these methods are suitable for keywords and image filtering only, hence we are proposing a method which recognize and classify the content that is adult text, adult image and adult video.

B. Proposed System-

Recognition and Classification of Adult Text, Adult Images, Adult Videos from Web Content

The proposed system works on Text, Images and Videos combining which presents a fuzzy ontology–based semantic knowledge system and support vectormachine (SVM) to systematically filter web contentand to identify and block access to pornography. For ill-suited content detection the fuzzy ontology provides semantic knowledge, and the SVM removes irrelevant content.

As web content consist of not only text but also include Images and Videos. We develop an integrated enhanced adult-content recognition system which can detect adult images, adult videos. In adult image-recognition algorithm, we model skin patches rather than skin pixels, resulting in better results than state-of-the-art algorithms which model skin pixels. In adult video-recognition algorithm, the adult video contains the information in which the audio section with an image is used to obtain a prior classification of the image. The algorithm achieves a better performance than the ones which use image information alone or audio information alone. The adult Web page recognition is carried out using multi-instance learning algorithm based on the combination of classifying texts, images and videos in Web pages. Both the speed and the accuracy for recognizing the Web

adult content are increased, in contrast to recognizing Web pages one-by-one.

IV. METHODOLOGY

This proposed system describes an integrated adult-content recognition system which can handle adult images, adult videos, and adult Web pages. The fuzzy ontology provides semantic knowledge for ill-suited content detection, and the SVM removes irrelevant contents. A predefined number of Web pages are selected from a Web page set that is composed of a Web page and the pages linked to it to form a Web page bag. We treat a Web page bag as a bag in multi-instance learning (MIL) and its Web pages are treated as instances in the bag. For each Web page in a bag, the different analysis strategies should be designed specifically to deal, respectively, with texts, images, and videos. We use text processing to extract text features of a Web page, we use the results of classifying images in the Web page to extract its images features, and we use the results of classifying its videos to extract its video features. Next, the text features, the image features, and the video features are concatenated to form the feature vector of the Web page. The feature vectors of the Web pages in the sample bags are used to construct the MIL-based classifier, which is used to classify test Web page bags.

V. CONCLUSIONS

We have developed a system which can recognize Web adult images, adult videos, and adult Web page bags. In our adult-image recognition algorithm, skin patches have been detected by modeling skin patches rather than skin pixels, and the recognition features have been extracted us-ing the idea of going from global to local. In our algorithm for recognizing adult videos, the result of recognizing the audio section associated with an image in a video has been used as a prior classification of the image. Our algorithm has achieved a better performance than the ones which use image information alone or audio information alone.

We have carried out recognition of adult Web page bags rather than individual Web pages. Bayesian-kNN citation-kNN has been used to recognize adult Web page bags. The idea of a fuzzy-ontology and SVM–based adult content detection system is proposed to automate the classification of pornographic versus medical websites. The results are very promising. It can also be used at home, in offices and schools, and in other public sectors to intelligently investigate a network. Furthermore, this system can overcome the classification problem with medical websites, since it can extract medical features from unclear webpage content, classify these features as either medical or adult, and calculate an indicator value for the decision-making system.

REFERENCES

[1] Farman Ali , Pervez Khan, KashifRiaz, DaehanKwak, Tamer Abuhmed, Daeyoung Park, Kyung Sup Kwak,"A Fuzzy Ontology and SVM–Based Web Content Classification System" in IEEE Communications Magazine, Volume 5, 2017, December 5, 2017, Digital Object Identifier 10.1109/ACCESS.2017.2768564.

[2] R. Du, R. Safavi-Naini and Willy Susilo, " Web fltering using text classifcation," in The 11th IEEE International Conference on Networks, 28 September - 1 October 2003, 325-330

[3] J. Wehrmann, G. S. Simıes, R. C. Barros, and V. F. Cavalcante, ''Adult content detection in videos with convolutional and recurrent neural networks,'' Neurocomputing, vol. 272, pp. 432–438, Jan. 2018.

[4] J. J. Sheu, ''Distinguishing medical Web pages from pornographic ones:An efficient pornography websites filtering method,'' IJ Netw. Secur., vol. 19, no. 5, pp. 839–850, 2017.

[5] M. Vanetti, E. Binaghi, B. Carminati, M. Carullo, and E. Ferrari, ''Content-based filtering in on-line social networks,'' in Privacy and Security Issues in Data Mining and Machine Learning (Lecture Notes in Computer Science). Berlin, Germnay: Springer, 2011, pp. 127–140.

[6] M. Wesam, A. Nabki, E. Fidalgo, E. Alegre, and I. De Paz, ''Classifying illegal activities on Tor network based on Web textual contents,'' in Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics, vol. 1. 2017, pp. 35–43.

[7] S. Seifollahi, I. Gondal, A. Bagirov, and R. Layton, ''Optimization based clustering algorithms for authorship analysis of phishing emails,'' Neural Process. Lett., vol. 46, no. 2, pp. 411–425, 2017.

[8] G. Xu, C. Wang, H. Yao, and Q. Qi, ''Research on Tibetan hot words, sensitive words tracking and public opinion classification,'' Cluster Comput.,2017. [Online]. Available: https://doi.org/10.1007/s10586-017-1026-x

[9] D. Yu, N. Chen, F. Jiang, B. Fu, and A. Qin, ''Constrained NMF-based semi-supervised learning for social media spammer detection,'' J. Knowl.- Based Syst., vol. 125, pp. 64–73, Jun. 2017.

[10] Z. Weinberg, M. Sharif, J. Szurdi, and N. Christin, ''Topics of controversy: An empirical analysis of Web censorship

lists,'' Proc. Privacy Enhancing Technol., vol. 217, no. 1, pp. 42–61, 2017.

[11] C H. Zheng, H. Liu, and M. Daoudi, ''Blocking objectionable images: Adult images and harmful symbols,'' in Proc. ICME, 2004, pp. 2–5.

[12] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank, ''Recognition of pornographic Web pages by classifying texts and images,'' IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 1019–1034, Jun. 2007.

[13] OuXinyu, Ling Hefei, Yu Han, Li Ping, ZouFuhao, Liu Si, "Adult Image and Video Recognition by a Deep Multicontext Network and Fine-to-Coarse Strategy," in ACM Transactions on Intelligent Systems and Technology, Vol. 8, No. 5, Article 68, Publication date: July 2017.

[14] Lee, P.Y. et al."Neural networks for web content filtering" in IEEE Intell. Syst. 17,

5, 48–57 (2002).

[15] J. Wehrmann, G. S. Simıes, R. C. Barros, and V. F. Cavalcante, ''Adult content detection in videos with convolutional and recurrent neural networks,'' Neurocomputing, vol. 272, pp. 432–438, Jan. 2018.

[16] R. Cohen-Almagor, ''Online child sex offenders: Challenges and countermeasures,'' Howard J. Criminal Justice, vol. 52, no. 2, pp. 190–215, 2013.