Distributed Agents for Web Content Filtering

Talib T. Al-Fatlawii
Department of Multimedia
College of Computer Science and Information Technology
University of AL_ Qadisiyah

Dr. Abbas M. AL_Bakery
University of Information Technology
And Communications

***Abstract*— This paper describe Web Content Filtering that aimed to block out offensive material by using Distributed Agents. The proposed system using FCM algorithm and other page's features (Title, Metadata , Warning Message) to classify the websites (using as candidate) into two types:- white that considered acceptable, and black that contain harmful material taking the English Pornographic websites as a case study.**

Keywords—Web Content Filtering; Fuzzy C-Means Algorithm; Agent and Multi- Agent System.

I. Introduction

In last recent year the (WWW) has become infinite information repository, and people became more dependent on Internet in their life, either for searching information, communication, e-commerce, e- mail,…,etc. But with this advantages, there exists some drawbacks, like offensive material that can be found on the websites (terrorism, violence, hate message, crime, pornographic material,…,etc), with reference to exist more than (4,200,000) websites in the world in 2013, i.e. 12% of the websites in world. Also 42.7% of internet users view the pornographic sites [1], A study in the southeastern U.S. found that 53 percent of boys and 28 percent of girls (ages 12-15) reported use of sexually explicit media. The Internet was the most popular forum for viewing. The words "sex" and "porn" rank fourth and sixth among the top ten most popular search terms[2]. This number may be increase in the next year, this type of sites is considered harmful for children and also for adult people, and could cause a side effect, therefore existing a system that can filtering these websites is necessary especially in home, school, university,…etc.

To address the problem of web content filtering system some strategies have been used, some used packet filtering approach, this method concern about IP address, but the IP address represents a particular host and this host can contain more than one sites, some of these sites considered acceptable, when blocking this IP this cause to block all the acceptable sites[3], also the control access list of IP is generated manually and this required great human efforts, Other used ***white/black lists*** of resources, they classified the sites to white and black, respectively. Such classification is performed by rating agencies, where manual collection and classification is required[4]. Other used ***Banned Word Lists***, this technique allows the creation of a blacklist dictionary that contains words or phrases. URLs and web content are compared against the blacklist to block unauthorized Websites. Vendors provide words blacklists with their products. And allowing the user to add new word to blacklist. The accuracy of this technique is considered un acceptable. for example, medical research sites are often blocked because they are mistaken for offensive material[5]. Other using AI technique such as ANN (Artificial Neural Network) to classify the sites into black and white list [6,7,8], this provide the ability to update these lists continuously and ensure the large number of undesirable sites are blocked.

In this paper we proposed a new approach for classification and blocking black (pornographic) sites. The system consists of distributed agents, its divided into two sides, the Server Side (Administrator Side) contain the following agents: the main agent is Classifier Agent where is responsible for classifying sites into two categories white (normal, acceptable) and black (pornographic) sites using FCM algorithm and some pages features, the result of this agent considered as candidate sites, the second agent called Administrator Agent this agent is responsible for generating black lists (consider very trusted classified sites), using the candidate sites generating by Classifier Agent, this agent deals with the administrator of the system. The third agent called Updating Agent, this agent charges of updating other agents that resides in each clients. In the Client Side there exist one agent called Filtering Agent, is responsible for blocking undesirable sites from reach the internet end users and taking update from server side.

This paper organized as follow: section1 the introduction, section2: Page Features, section3 explains the proposed systems, section4 explain the implementation, and finally section5: Conclusion and Experimental results.

II. Page Feature for Classification

The page features are important in each classification process, some of these features we used here are :

1) *Title: Represent the the title of webpage, it is easily to deal with where it found between <TITLE> and*

</TITLE> tags in HTML page, the title of page represent it's subject [6].

2) *Meta Data: Also considered important componenet in page, where it contains information about the WebPage, for example Meta Description contains a short description about the web page, it allows the developer to summarize the content that can be found on the page. Also Meta Keyword contains the keywords that provide short and accurate information about the web page [9,10].*

3) *Warning Message Box : The warnings message block is used to alert visitors to the explicit contents of the Web site and the legal aspects involved. Is used to relieve any legal responsibility resulted from visitors viewing the pornographic contents. It contains number of legal terms like " Warning This Site Contains Adult Content", " sexually explicit material", "This website is for adults only" and so on [6].*

III. The Fuzzy C – Means

Is one of the most important and popular unsupervised partitioning algorithm used in several application domains such as pattern recognition, machine learning and data mining, feature analysis, clustering and classifier design…etc[11,12].

The clusters are formed according to the distance between data points and cluster centers are generated for each cluster. Where each points (x_i) can belong to more cluster in same time, but with degree of membership according to distance between the cluster center (v_j)and the point. It is based on minimize the following objective function

$$J_m = \sum_{j=1}^{c} \sum_{i=1}^{n} \mu_{ij} d_{ij} \qquad (1)$$

Where $$d_{ij} = \left\| x_i - v_j \right\| \qquad (2)$$

The Algorithm Fuzzy C-Means steps are[12]:

1. Initialize U=[U_{ij}] matrix, $U^{(0).}$
2. At k – step: calculate the centers vectors $C^{(k)}=[c_j]$ with $U^{(k)}$.

$$v_j = \frac{\sum_{i=1}^{n} \mu_{ij}^{m} x_i}{\sum_{i=1}^{n} \mu_{ij}^{m}} \qquad (3)$$

3. Update $U_{(k)}$, $U_{(k+1)}$.

$$\mu_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{d_{ij}}{d_{ik}} \right)^{\frac{2}{m-1}}} \qquad (4)$$

4. If $\| U_{(k+1)} - U_{(k)} \| < \varepsilon$ then Stop, otherwise return to Step 2.

IV. The Proposed System

Our system based on distributed agents that coordinate with each other in order to perform the task of web content filtering system. Figure (1) shows the structure of the system. In this structure we show two sides, the first one is the server side which is charge of classifying sites and updating the client agent. The second side is client side, which is charge of blocking out undesirable websites, and also taking update from server side.

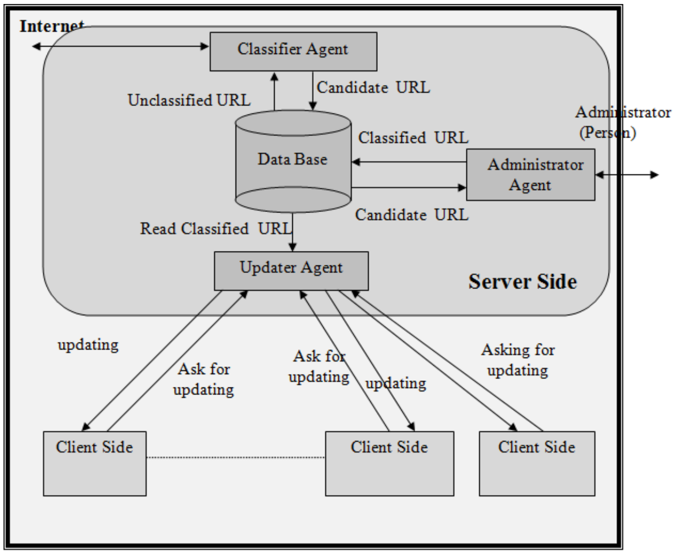


Fig. (1) shows the architecture of the System

The system consists of the following stages:

A. *Impelementation of FCM algorithm*

First, we must compute the cluster center for each sites' category in order that Classifier Agent using it in the classification process. In the proposed system we have used the textual features of page, taking 100 pages in the implementation process, converting it to vectors, where each vector contain numbers that represent count of the words or phrases that frequently appeared in pornographic pages. Here we used 45 words and phrase, after gaining the 100 vectors we applying FCM algorithm to get two clusters centers, one represents the white (normal) and the other represents the black (pornographic) pages. These centers later used in the classification process by the Classifier Agent.

B. *Classifier Agent*

This agent is responsible for classifying the sites into white and black categories, using some page features like [Title, Metadata, and Warning Message] and FCM, figure(2) shows the architecture of Classifier Agent. This agent represent the core of the system. It is used the clusters centers that computed from the previous stage in the classification process. And after this store the result as candidate in the database where it is can be changed by the administrator of the system.

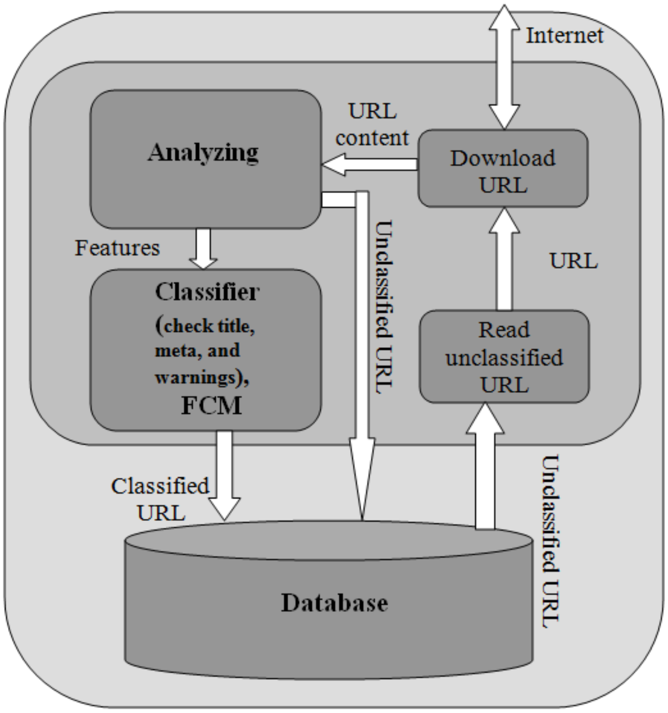
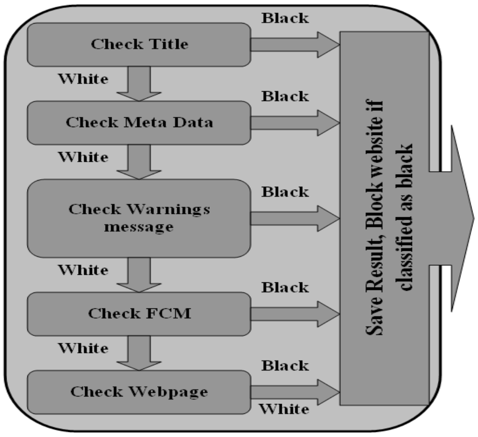


Fig.(2) shows the Classifier Agent

The classifier model in the figure above consist of five stages, figure (3) shows these stages.



The Classifier model

The ***Check Title*** and Check ***Meta Data*** stages search for specific words and phrases in the title and meta data of page, if it is found it, then classifying the sites as a whole as a black sites. Also the ***Check Warnings Message,*** searching for specific statement, this statement used to alert users to content of the website, Is used to relieve any legal responsibility resulted from visitors viewing the pornographic contents. It contains number of legal terms like "Warning This Site Contains Adult Content", "sexually explicit material", "This website is for adults only" and so on, if it found these statements, then considered the sites as black sites. Where the ***Check FCM*** classify the page based on converting page to vector (the vector contain the number of words and phrases that frequently appeared in black sites), and compute the membership to each cluster center, later if the page that have membership tend to white cluster center then considered the page as white, else considered as black. Later the final stage is the ***Check Webpage***, in this stage the classifier check all the pages that linked to current page, using the functions (Check Title, Check Meta, Check warning, Check FCM) that explained previously, if found two pages that classified as black, consider the site as black, otherwise it is considered white.

This agent also working as web pages crawler, when it downloads a specific page, it analyzing it, and extracting webpages and sites that connected to it. Later these sites are classified too.

C. *Administrator Agent*

This Agent deals directly with the administrator of the system. It explain all the sites that found in database, it shows the sites that classified by the Classifier Agent, allowing the administrator to ensure its type or change in order that the results that send the client side are very trusty.

D. *Updating Agent*

This agent is responsible for updating the database of system, it is using the sockets to exchange the message with the Filtering Agent, it continuously receive the update's request that send by the Filtering Agent and sending the last websites that classified as black in order to block it. It is used java socket, to exchange message between agents.

E. *Filtering Agent*

This Agent is responsible for filtering unaccepted websites. This agent using JPCAP package[14] for capturing every packet that sending by the user, if found one packet that goes to undesirable site then killing the browser (closing the browser), as we mentioned previously this agent's database continuously updated by the Updating Agent, and this provide ability The Filtering Agent working as Anti-Porn program, or parental control software that reside on each client. Figure below shows the FA.

V. IMPLEMENTATION

The system was implemented on network consisting of 3 PC, using Java language for constructing the agents, JPCAP for capturing packets, JSOUP package to download and parsing the Page's HTML code[15], MySQL language for constructing Database. Using FIPA ACL message for the communication between server side agents and client side agent.

VI. CONCLUSION AND EXPERIMENTAL RESUTLS

Design web content filtering system using distributed agent could improve the efficiency of the system as a whole where agent have properties like autonomous, social, flexible…etc. This properties are necessary in such system. Also using agent in the update of system database will increase the performance of the web content filtering system and ensure that maximum number of pornographic sites will blocked.

Using the classified websites as a candidate where is later checked by the administrator of the system, this make the blocked sites is trusty. And also making the system more efficient when it is compared to traditional web content filter because it is blocked the sites after checked by the administrator of system where the error is removed

REFERENCES

[1] URL, "http://internet-filter-review.toptenreviews.com/internet-pornography-statistics.html", date of access 2013.

[2] URL," http://www.internetsafety101.org/Pornographystatistics.html", date of access 2014.

[3] A.C.M. Fong, et al. "An intelligent offline filtering agent for website analysis and content rating", 2nd Symposium on Web Society (SWS), IEEE , pp. 1 – 4, 2010.

[4] Elisa Bertino, et al. "A General Framework for Web Content Filtering", World Wide Web, Springer, Vol 13, Issue 3, pp. 215 – 249, September 2010.

[5] John R. Vacca, "Computer and Information Security Handbook", Elsevier Inc, 2009.

[6] P. Y. Lee, et al. "An Intelligent Categorization Engine for Bilingual Web Content Filtering", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 7, NO. 6, pp. 1183 – 1190, 2005.

[7] Ali Ahmadi, et al. "Intelligent classification of web pages using contextual and visual features", Applied Soft Computing, Elsevier, Vol 11, Issue 2, pp. 1638–1647, March 2011.

[8] Akhan Akbulut, et al. "Agent Based Pornography Filtering System", International Symposium on Innovations in Intelligent Systems and Applications (INISTA), IEEE, pp. 1 – 5, 2012.

[9] URl, "http://www.tizag.com/htmlT/meta.php", Date of access 28/3/2013.

[10] Lim Wern Han and Saadat M. Alhashmi, 2010, "Joint Web-Feature (JFEAT): A Novel Web Page Classification Framework", IBIMA Publishing, Communications of the IBIMA, http://www.ibimapublishing.com/journals/CIBIMA/cibima.html, Vol. 2010 (2010), Article ID 73408, 8 pages, DOI:10.5171/2010.734081.

[11] Guojun Gan, et al. "Data Clustering Theory, Algorithms, and Applications", American Statistical Association and The Society for Industrial and Applied Mathematics, 2007.

[12] Bruno A. Pimentel, Renata M.C.R. de Souza, "A multivariate fuzzy cmeans method", Applied Soft Computing, Elsevier B.V., Vol 13, Issue4, pp. 1592–1607, April 2013.

[13] URL,"http://home.dei.polimi.it/matteucc/Clustering/tutorialhtml/cmeans.html", Date of access 7/3/2013.

[14] URL," http://www.eden.rutgers.edu/~muscarim/jpcap/index.html", Date of access 2013.

[15] URL, "http://jsoup.org/", Date of access 2013.