Module of Text Information Analyze in the Personified System for Information Filtering

Roman Zharinov, Ulia Trifonova, Alexandr Kodyakov, Oleg Karmaleev
Saint-Petersburg State University of Aerospace Instrumentation
Saint-Petersburg, Russia
{roman, ulia}@vu.spb.ru, kod.aleksandr@gmail.com, axelk21@yandex.ru

Abstract

Text analyze for future filtering is one of the most useful tasks for children education, because it's hard to imagine modern education in schools, without usage of Internet resources. This work is aimed at techniques and technologies of filtering text information that can harm to children.

Index Terms: DLP, Text Analyze, ICAP-server.

I. Introduction

Nowadays modern education, especially at schools, cannot dispense without using Internet, both at home and at schools or some educational buildings. However, the use of Internet resources, is out of control, leading the necessity filter the potential dangerous content, that is necessary to have ability to disable or limit access to types of information specified in the federal law from 29.12.2010 № 436-FZ [1], [2]. There are at least two reasons causing the creation of a filter system [3]:

- Prevent receiving information that could cause harm to children.
- Prevent a possibility of infection by the computer virus, there is the possibility of loss or leakage of critical or sensitive information (both located on the same computer in a local network).

In Russian Federation, this issue is especially relevant because within the priority national project "Education" planned to be provided access to Internet resources by more than 10,000 schools in the country. It’s noticed the imperfection and low efficiency the filter systems discussed at the round table on the issue of the child safety on the Internet [4]. In 2012, it proposed to implement a system of personalized filtering the age group of users [1], [5].

II. Main part

Wide use of modern techniques of designing client-server systems “Web 2.0” rather complicated content-filtering Internet traffic, i.e. in most cases, data is transmitted separately from the design, thus there is possibility to skip unwanted information as from the user and to him (income and outcome traffic respectively). In the case of that method of the design, it is necessary to conduct a comprehensive analysis of data transmitted during the whole “session” user with a client-server system.

As a core of the collection and processing of security incidents we use the system to prevent leaks of confidential information MyDLP [5], [6]. To intercept the incoming information we use protocol ICAP (Internet Content Adaptation Protocol) [7], designed to implement the service content filtering features online translation embedding/cutting out ads, etc.

ICAP is a lightweight HTTP like protocol which is used to extend transparent proxy servers. For content analyzing this protocol has a few added services (content manipulation) for the associated client request/response. In the role of ICAP-client using any system, through which can go all network traffic (usually a similar system is a proxy server). ICAP-server can function as clients for other ICAP-servers that allow creating a collective processing of data transmitted (see Fig. 1). This protocol allows modifying all the queries and replies of ICAP-server.

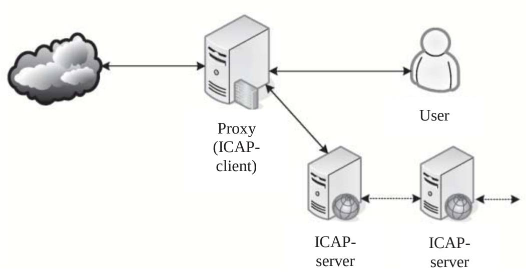

Fig. 1. ICAP-server in a client role

Using that protocol we can produce any number of operations such as [8]:

- Content substitution – received content may be replaced by some new content, e.g.:
  - Automated translation to needed language.
  - Remove banner ads by filtering on size, some templates, etc.
- Re-coding or modification markup of content – transform content from “human view” to best view for analyses
  - HTML type to WAP/XML – translate from html document to plain text in specific format.
  - Bmp type of image convert to png/gif – reduce size to the client side.
- Access control – grand or deny access to the internet or some resources, e.g.:
  - Authorization procedure.
  - Providing access to work or non-work content.
- Virus scanner – one of the most popular and useful task for this protocol.
- Content compression – turn on gzip compression for all text traffic.

Disadvantages of using ICAP-protocol in the network infrastructure [9]:

- Additional delay in the network between the client and the server – speed of transfer data may decrease between external systems and data consumers.
- Perform additional checks (data type definition, size, etc.) of the ICAP-server. This is due to the fact that in many cases, ICAP-clients use the file extension to determine the type of data.
- Difficulty in integrating with systems using protocols other than HTTP, thus not allowing using the ICAP-depth analysis of a protocol for data.

There are four main operations that support and provide ICAP-protocol:

- Request Modification (Fig.2)
  - Client sends request to proxy server (server which client will be receiving content).
  - Proxy server forwards request to the ICAP server

- ICAP server will respond with some modified content or headers.
- Proxy server will use that modified request header to process the request.

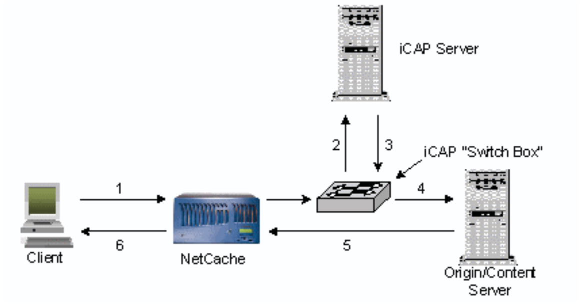

Fig. 2. Request Modification

Main application: filtering headers, redirection to deny pages, etc.

- Request Satisfaction (Fig.3)

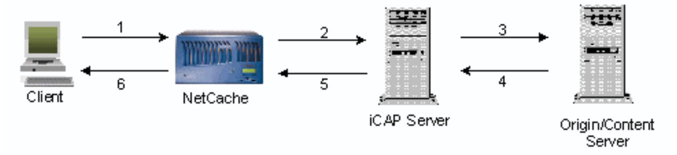

Fig. 3. Request Satisfaction

- Client sends request to proxy server.
- Proxy server forwards request to ICAP server.
- ICAP server will respond with a modify query.
- Proxy server will pass that response header and body onto the client.

The request will not be further processed.

Main application: blocking domains, sites by IP-address or by URL

- Response Modification (Fig.4)

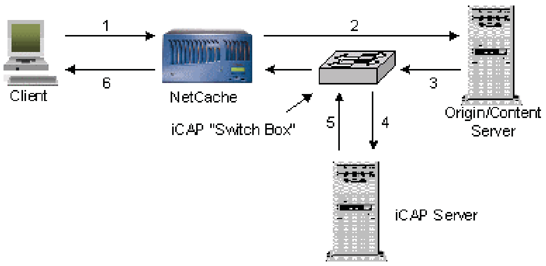

Fig. 4. Response Modification

- Client sends request to proxy server.

  - Proxy requests content from web server than forwards response header and body to ICAP- server.
  - ICAP server will respond with a possibly modified response header and body.
  - Proxy server will then send the possibly modified response header and body to the client.

Main application: content modification, virus scanning, block inappropriate content, etc.

- Result Modification (Fig.5)
  - Client sends request to proxy server.
  - Proxy server forwards request to ICAP server with a possibly modified response header and body.
  - After that, ICAP forwards response to Proxy server.

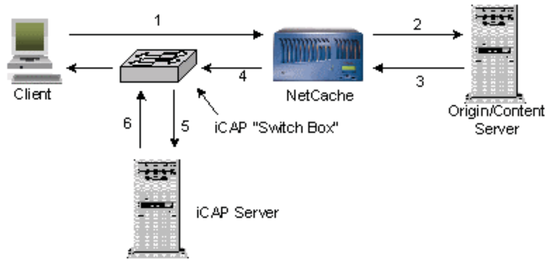

Fig. 5. Result Modification

As that type of ICAP-server situate clothes to client, proxy server can cache some client content. When an ICAP server is installed with a caching system, every transaction is piped through the ICAP server, allowing the server to modify or redirect Web requests or responses.

In general, the proxy server works as ICAP-client. So, when user interacts with the destination server through a proxy, support of ICAP-protocol does not need for the end user, and traffic adaptation for him looks like “transparent”.

The preferred mode of operation for us when ICAP-server can modify the results, as it allows us to analyze the cached objects (images, video, etc.).

The following is a list of the existing ICAP solutions as well as toll-free to use – Table I.

Develop a system of personalized data filtering can be classified as independent systems which install and configure within individual local networks or organizations.

A module allows analyzing information on the following grounds:

- General (meta-) information about the resource.
- Content of the resource.

Analysis of the information occurs in real-time (on-line), ie during the response from the Internet resource user.

The main quantitative indicators of similar systems filter information (mainly text-filtering) are:

- The accuracy of the analysis - the percentage of correctly blocked Internet resources.
- Type I error (False positive error) – excessive filtering "good" resources.
- Type II error (False negative error) – insufficient filtering or false negative response content of the web resource.

TABLE I
ICAP SOLUTIONS

	Programming language	Licenses	Main application	Cost
C-ICAP	C++	GNU GPL v2, GNU LGPL v2.1	Checking network traffic for viruses	Open Source
ICAP-server	Python	GNU GPL v2, CNRI Python License	Translate pages from English to French	Open Source
POESIA	Java	GNU GPL	Text filter for the five languages	Open Source
GreasySpoon	Java	Affero GNU Public License	Allows to write addons in different languages , high performance	Open Source
Customised ICAP Server	C	-	Content filtering, virus scanning, translation and more	Open Source
WebFlow Adapter	Java	-	To bind HTTP analysis to your legacy authentication system. To filter traffic using complex rules	290€

Currently, the total error of algorithms in the system to dynamically filtering Internet resources remains low:

- At highest practicable accuracy of the analysis 90% of the systems have either a very high percentage of false positives (about 2-5%) which leads to blockages about every 20th page.
- Or low speed of analysis, which causes significant delays on the client side (more than few seconds).

At the moment text filtering module is implemented as a separate module for the system to counter the leakage of confidential information MyDLP using the Squid proxy and ICAP-server, allowing caching and modify information. This architecture allows one to easily manage the security policies and minimally dependent on the network infrastructure of the organization. Developed a scheme for using the module is shown in Fig. 2.

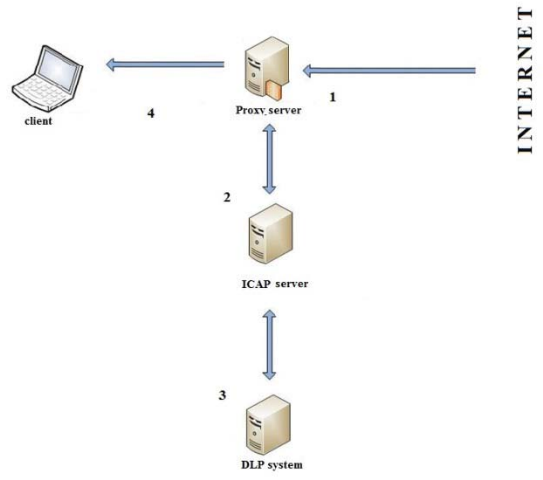

Fig. 6. The interaction of the components

Algorithm of work:

1) Intercepts incoming flow of information an Internet resource with helping proxy server Squid and redirect it to the ICAP-server.
2) Preliminary analysis of content:
   - Classification of the analyzed information (media content or text content).
   - Analysis of information on potentially dangerous words on the treated Internet resource and further redirect the system MyDLP.
3) Delivery or blocking the requested content of Internet resource to the client.

Traditionally, in the filtering systems of Internet resources used signature search, using database of addresses – pair of key-value consisting of the domain name and the category of Internet resource. However, such systems have a number of disadvantages, for example, the analysis of page content in real-time – i.e. the ability to dynamically change content in a trusted online resource.

Thus, the use of the signature-based scanning systems for traffic analysis does not allow adapting to the constant changes in the dynamics of content on the Internet resources. That's why developed module must have the ability to automatically adapt to the dynamically changing content of the current site without taking into account the category of the selected Internet resources. Also plans to implement the mechanism of decision-making, which can use any kind of information about Internet resources and users, in particular, it may request additional information from the mechanism of classification (the site). If the Internet site was previously analyzed, decision-making mechanism can request results from the cache classification, hash of content present in the cache classification.

III. Conclusion

Thus, the filter module combines both text information content filtering of Internet resource:

- Dynamic filtering – analysis of the contents in real-time.
- Domain filtering – based on permissive and deny lists of domain names of Internet resources.

The need to apply both types of filtration caused by the fact that as a content analysis algorithm of Internet resources using frequency analysis of the occurrence of potentially dangerous words in the security policy. Thus, to determine the subjects of a site you can with a certain probability. The need to apply both types of filtration caused by the fact that as a content analysis algorithm of Internet resources using frequency analysis of the occurrence of potentially dangerous words in the security policy. Thus, to determine the subjects of a site you can with a certain probability.

The procedure for customizing templates for dynamic filtering, to avoid the error is quite long and is always performed by the developer, or security administrators. In turn, the use of "black" and "white" lists of Internet resources is given an opportunity to make adjustments in the content filtering web pages.

References

[1] Federal law of 29.12.2010 № 436-FZ "O zashchite detey ot informatsii, prichinyayushchey vred ikh zdorovyu i razvitiyu" (in Russian).

[2] Zharinov Roman, Trifonova Ulia, Concept of the system to protect children's access to information in education institutes using RFID-technology, *Scientific and Technical of Information Technologies, Mechanics and Optics*, 2013 (in review process, in Russia).

[3] K.V. Moiseev, Dynamic method of filtering Internet sites with aggressive content.

[4] "Rostelecom" held "round table" on the issue of child safety on the Internet, Web: http://www.voronezh.center.rt.ru/press/news/news2678.

[5] Trifonova, U. V. и Zharinov, R. F., Concept of the System to Protect Children’s Access to Information, *Proceedings of the 12th Conference of Open Innovations Association FRUCT*, 2012, pp. 142-146.

[6] MyDLP - Data Leak Prevention Solution. Official resource, Web: http://www.mydlp.com/.

[7] Network Working Group, Internet Content Adaptation Protocol (ICAP), Web: http://tools.ietf.org/html/rfc3507.

[8] Enno Davids, ICAP - The Internet Content Adaption Protocol, AUUG 2004 - *Who Are You?*

[9] Modern trends in content filtering, Web: http://alexott.net/ru/writings/cf/index.html.

[10] U.V. Trifonova, “Kids. Protection from ineligible content”, *XIII International Forum Modern information society formation - problems, perspectives, innovation approaches: Proceedings of the International Forum*, Saint-Petersburg, SPb.: SUAI, 2012, pp. 186-190.