Selection of countermeasures against propagation of harmful information via Internet

L A Vitkova, A P Pronichev, E V Doynikova, I B Saenko
SPIIRAS, liniya 14-ya, 39, St.Petersburg, Russia

E-mail: vitkova@comsec.spb.ru

Abstract. Today, Internet allows accessing and propagation various information. Some of this information can be undesired for particular user or even harmful. While there are quite a large number of systems to respond to unwanted and harmful information and some well-established set of countermeasures, such as filtering, blocking access, and notification, existing systems are often based on one type of countermeasures. From our point of view, there is a need for the technique of automated selection of optimal countermeasures to counteract undesired or harmful information on the Internet depending on the type of such information and characteristics and needs of the protected system. In this paper we propose the set of interconnected models, including threat model, countermeasure model and information object model, and countermeasure selection technique for protection against harmful information in Internet using these models. The proposed technique uses single-criteria optimization on the basis of introduced countermeasure selection index and allows selecting applicable and optimal countermeasures for the particular system from specific threats. The application of the technique is demonstrated on the experiments.

1. Introduction
Today, Internet allows accessing and propagation various information. Some of this information can be undesired for particular user or even harmful. Harmful information can include annoying suggestions for gambling or lottery, a proposal, request and/or order to commit suicide, and/or description of suicide as a way to solve problems, data about production, transfer, sale or storage of child pornography, description of the production, development and use of drugs and psychotropic substances, and others. Some information can be forbidden for propagation by legislative acts, other can be specified as undesired by particular user. Today, parental control systems filter traffic based on security policies. Social media platforms provide protection from unwanted information for a child if the user has specified their age. This protection is also provided by some operating systems, such as iOS, by blocking access to pages with such content. Again, the user must specify the correct age. When implementing parental control security policies in a social network or operating system, the systems themselves only analyze messages and pages that the user interacts with. The tasks of countering the propagation of unwanted and harmful information, its distributor, and evaluating user behavior on such pages are not set or solved by social network platforms.

In fact, all existing tools implement the following actions: 1) checking information, 2) restricting access to a resource with unwanted information [1, 2]. While today there are quite a large number of systems to respond to unwanted information and some well-established set of countermeasures, such as filtering, blocking access, and notification, existing systems are often limited with above mentioned approach based on one type of countermeasures, i.e. blocking access. From our point of view, there is
a need to classify countermeasures, and develop the models and technique for automated selection of optimal countermeasures to counteract unwanted or harmful information on the Internet depending on the type of such information and characteristics and needs of the protected system. In this paper we propose the set of interconnected models, including threat model, countermeasure model and information object (IO) model, and countermeasure selection technique for protection against harmful information in Internet. The proposed technique uses single-criteria optimization on the basis of introduced countermeasure selection index and allows selecting applicable and optimal countermeasures for the particular system from specific threats.

The paper is organized as follows. The second section provides the results of the related works and solutions research. The proposed models and technique are given in Section 3. The implementation and experiments are described in Section 4. The paper ends with conclusions and discussion about future research.

2. Related works

Parental control systems, anti-virus programs and methods, models and algorithms developed in this field are the first to protect against unwanted information.

Due to the huge volume of media content available from various sources (for example, TV, Internet, online games, etc.), parental controls are becoming increasingly important for censorship and access control of media content. Parental control systems traditionally allow users to restrict access to specific types of media content based on media characteristics, such as ratings, time, or resource type. In [3], mechanisms for filtering access to media content are proposed. The authors [4] in their work suggest a way to intercept information about the file at the moment when the file is saved to the device and can be run in single-user mode (offline). In [5], a method for controlling traffic and access to information in a network environment in a local network is proposed. In [6], a method for making decisions about parental control approval for user content is proposed, but with the parent's approval of the child's rights to upload a media file from an external website or resource.

Most existing systems for countering unwanted information on the Internet use methods based on the classification of web pages, resources, and the use of countermeasures such as lists (url blacklist). However, traditional methods are improved by using machine learning algorithms. In [7] the authors state that the blacklist is now completely ineffective in searching for both malicious URL variants and newly created URL's. In addition, creating and updating lists requires human input. Blacklisting is a time-consuming process. Machine learning solutions implicitly rely on the function development stage to extract characteristics. Such as: linguistic, lexical, contextual, or semantic attributes, URL string statistics, n-grams, word packets, link structures, content composition, DNS information, network traffic, etc. As a result, machine learning solutions must recognize new malicious URL's, and these systems are not yet accurate enough. The authors tested few deep learning methods such as recurrent neural networks, long-short-term memory, convolution neural networks and identity-recurrent neural networks. Algorithms and methods based on deep learning can extract objects automatically by accepting raw input texts. In the [8] shows a method for detecting malicious web pages by analyzing the domain and scanning it.

In fact, the black and white list policy is not really applicable to protect against unwanted information, for example in social networks, because the URL's of an IO contain not the attributes of an IO, but the attributes of the path of the object. The domain of the social network platform, the authors in it, and the messages are connected on the page of news portal. Adding the path of such a portal to the blacklist leads to excessive blocking.

The need to use algorithms for semantic text analysis leads owners of systems for countering unwanted information to the problem of analyzing extremely large amounts of information. Search and analytical systems and their clients (browsers) take the lead in developing such approaches. In the [9], a machine learning method is demonstrated that allows us to study a variety of messages in the social network Twitter, user moods, and conduct analytics. Such works are also popular in the field of marketing, economics and finance to assess consumer sentiment. In the [10] a method is proposed for evaluating the tone of messages in the social network Twitter, based only on smiles. In the [11], the Heracles platform for developing text analysis algorithms is proposed.
In addition, there are also many solutions that protect users of the Internet and social networks from unwanted information based on the analysis of the graph of relationships. Patent [12] is devoted methods for protecting a social network user from prohibited objects (text, photo, media file) that were previously saved in the database. This technical result is achieved by building a social graph, a cluster of accounts in a social network, and a database of prohibited objects that are associated with social graphs and accounts. The invention relates to parental control systems that restrict user access to social network resources in accordance with the rules set by the parent. In [13], a method for quickly identifying similar communities and social network participants is proposed. During the analysis process, the system selection criteria for similarity of profiles and generates a list of IDs accounts that meet these criteria. Finally, in the created social graph, edges characterize the direct or indirect connection between social network accounts.

A review of existing studies shows that most of the existing solutions are based on the methodology within which the IO is checked. If unwanted information is detected, access to the resource will be limited. Also, a review of relevant studies shows that today there is some well-established set of countermeasures, such as filtering, blocking access, and notification. However, existing systems for countering unwanted information use only one type of measure from the established set of countermeasures. Therefore, there is a need to classify countermeasures, and develop a model and methodology for automatically selecting optimal countermeasures to counteract unwanted information on the Internet.

3. Countermeasure selection technique

The developed technique is based on the analytical modelling of counteraction process and decision support methods. In process of analytical modelling we specified the object affected by the harmful information, the threats against this object, and countermeasures against these threats. The developed technique uses single-objective optimization for countermeasure selection index. The first subsection considers the countermeasures and their types, and their applicability against particular threats, as well as the models of countermeasure, threat and object. The second subsection describes the proposed technique.

Countermeasures and models for countermeasure selection

We understand the countermeasure as an action or tool for protection of object from information security threat, where information security threat is represented by harmful information. We limit object with IO in Internet. We outlined the following properties for information objects specification: object size; type of information object (harmful or not); and harmful object class. Based on the outlined properties we specified the object model as follows: $IO = \langle \text{size}, \text{type}, \text{class} \rangle$, where size – is a class of information object, it can take values $\{s, m, l\}$, where $s$ – small (sio), $m$ – medium (mio), $l$ – large (lio); type – is a type of information object, it can take values $\{h, n\}$, where $h$ – harmful objects, $n$ – not harmful objects; class – is a class of harmful information object, it can take values $\{\text{cour}, \text{reg}, \text{lot}, \text{suic}, \text{por}, \text{dru}, \text{from}_6, \text{from}_{12}, \text{from}_{16}, \text{from}_{18}, \text{none}\}$, where cour – IO containing data earlier forbidden by court's decision; reg – IO containing data earlier blocked by authorized bodies; lot – IO containing suggestion for gambling or lottery; suic – IO containing proposal, request and/or order to commit suicide, and/or describing suicide as a way to solve problems, as well as endorsement of suicide; por – IO containing data about production, transfer, sale or storage of child pornography, the acquisition of child pornography; dru – objects containing description of the production, development and use of drugs and psychotropic substances and/or description of the cultivation locations of plants containing drugs and/or places of their sale and prices; from6 – IO containing information for children from 6 years old; from12 – IO containing information for children from 12 years old; from16 – IO containing information for children from 16 years old; from18 – IO containing information forbidden for children; none – not applicable (if type is n). The class of affected threats connects threat model and countermeasure model, while class of the information object connects countermeasure model and object model. Connections between the threats, countermeasures and object size are shown in table 1 on the example of drug information threat.
We analyzed possible countermeasures and outlined the following properties for their specification: the class of affected threats; implementation type; implementation agent; class of the information object; and complexity of implementation. Based on the outlined properties we specified the countermeasure model $rm$ as follows: $rm = <rm\_name, rm\_thr, rm\_obj, rm\_impl, rm\_agent, rm\_cost>$, where $rm\_name$ – countermeasure type (message filtration for sio in browser; block mio/lio URL by internet provider; mio/lio URL filtration by antivirus/parental controls/browser/operation system; black SEO; user notification for sio/mio/lio by browser/antivirus/parental controls/operation system; administrator notification for sio/mio/lio by browser/antivirus/parental controls/operation system; internet provider notification for sio/mio/lio); $rm\_thr$ – the class of affected threats (cour, reg, lot, suic, por, dru, from6, from12, from16, from18, none); $rm\_obj$ – class of the information object (small, medium, or large); $rm\_impl$ – implementation type (manual or automated); $rm\_agent$ – implementation agent (browser, internet provider, parental controls, antivirus, mobile application, operation system); $rm\_cost$ – complexity of countermeasure implementation (it depends on $rm\_impl$ and $rm\_agent$ and currently specified by the experts). The class of affected threats connects threat model and countermeasure model, the class of the information object connects countermeasure model and object’s model, and agent and implementation type connect countermeasure model with protected information system. Connections between the threats, countermeasures and object size is provided in table 1 on the example of drug information threats.

| Threat class | Threat description | IO size | Countermeasure |
|--------------|-------------------|--------|----------------|
| Drugs (Th4)  | Information object containing description of the manufacture, development and use of drugs and psychotropic substances and/or description of the cultivation locations of plants containing drugs and/or places of their sale and prices | SIO    | Message filtration for SIO in browser |
|              |                   |        | User notification for SIO by browser |
|              |                   |        | User notification for SIO by antivirus |
|              |                   |        | User notification for SIO by parental controls |
|              |                   |        | User notification for SIO by operation system |
|              |                   |        | Administrator notification for SIO |
|              |                   | MIO    | Block MIO URL by internet provider |
|              |                   |        | MIO URL filtration by antivirus |
|              |                   |        | MIO URL filtration by parental controls |
|              |                   | LIO    | Block LIO URL by internet provider |
|              |                   |        | LIO URL filtration by browser |
|              |                   |        | LIO URL by parental controls |
|              |                   |        | LIO URL filtration by antivirus |

The countermeasures should stop propagation of harmful information. The technique for countermeasure selection is provided in the next subsection.

**Technique for countermeasure selection**

The developed technique gets as input the harmful information objects, the threats relevant for this object and the list of countermeasures specified using the models described in the subsection below. The technique includes two main stages: (1) selection of the applicable countermeasures considering the protected system, information object’s and threats’ characteristics; (2) selection of optimal countermeasures on the basis of integral countermeasure selection index. The technique incorporates the following steps:

1. Input information object $io$ that have $type=h$, $class$ is given as input from harmful information detection system.
2. Get $rm\_agent$ and other characteristics that depend on the protected system.
3. Determine the set of possible countermeasures $cms$ for $io$ considering $class$, $size$ and $agent$, and characteristics of the protected system.
4. For each countermeasure $c$ from $cms$:
   - Determine countermeasure selection index $rm\_ind$ as integral sum of weighted countermeasure parameters’ values including implementation agent and implementation type (see countermeasure model above).
5. Select countermeasures \( scms \) from \( cms \) with max \( rm\_ind \).

6. Output \( scms \).

As the result the technique outputs the set of optimal countermeasures against detected harmful information object for the particular system under protection.

4. Implementation and Experiments

The developed technique was implemented using Java 11. For the experiments the dataset containing 10000000 objects of different size (sio - 33,33%, mio - 33,34%, lio - 33,32%) containing different threats (reg – 14,9%, lot – 14,89%, suic – 59,51%, por - 14,89%, dru – 29,76%, from18 – 14,87%, other 44,67 – see section 3). The input data set used in the experiment is a file containing the strings represented by a json objects with the following structure: (a) "SIZE" – the size of the IO; (b) "id" – the unique identifier of the IO; (c) "threats": ["t1", "t2", "t3"...] – an array of threats contained in the object. In the output set, an array of countermeasures is added to the existing structure, consisting of the following objects: (a) "name" – the name of the countermeasure; (b) "complexity" – complexity of countermeasure implementation; (c) "priority" – priority of the threat; (d) "type" – type of implementation (automatic, manual, etc.); (e) "agents": ["a1","a2"...] – array of agents to implement the countermeasure. For example, in case of input: "SIZE"="MEDIUM"; "threats"= "suicide"; and the system with characteristics: "agents"="ANTIVIRUS"; the output set will be: "countermeasures": "name"="FilterAntivirus","complexity"=3,"priority"=1,"type"="AUTO","agents"="ANTIVIRUS".

The analysis of the countermeasure selection component results allowed obtaining the following statistics: (1) The set of possible countermeasures included: (a) 3 countermeasures for 7,28% of information objects; (b) 4 countermeasures for 92,72% of information objects. (2) The countermeasures added to the set of possible countermeasures distributes as follows: (a) message filtration in browser - 31,72%; (b) filtration by antivirus - 44,37%; (c) filtration by parental controls - 0,87%; (d) administrator notification - 100,00%; (e) internet provider notification - 100,00%; (f) block by internet provider - 45,35%; (g) other - 14,03%. (3) The countermeasures added to the set of optimal countermeasures distributes as follows: (a) message filtration in browser - 31,72%; (b) filtration by antivirus - 44,37%; (c) filtration by parental controls - 0,87%; (d) administrator notification - 23,04%.

In figure 1 the distribution of the technique operation results, i.e. distribution of the optimal countermeasures against simulated information objects containing harmful information (threats), is represented.

![Figure 1](image)

**Figure 1.** Distribution of optimal countermeasures for the system modelled in the experiment. Where 44.37% - the countermeasures implemented by antivirus filtering tools; 31.72% - filtering in the browser; 23.04% - notification to the resource administrator where the information object is located; 0.86% - filtering by the parental control system.

As part of the experiment, we used a set of threats, information signs, and objects specific to the legal field of Russia. However, any operator, whether a parent or an authorized body, can set its own rules and countermeasures. The proposed technique is flexible (can be implemented on the user’s/internet provider’s/authorized body’s side) and automatically determines what countermeasures are the most or least available and optimal in the given conditions.

5. Conclusion

The paper considered the challenge of counteraction against propagation of undesired for particular user or even harmful information via Internet. The models of counteraction process, including the threat model, countermeasure model and information object model were developed. The technique for
automated selection of optimal countermeasures to counteract against such information using the developed models is proposed. It considers the information type and characteristics and needs of the protected system. It uses single-criteria optimization on the basis of introduced countermeasure selection index and allows selecting applicable and optimal countermeasures for the particular system from specific threats. The technique is flexible and can be implemented on the user’s, internet provider’s or authorized body’s side. The operation of the technique was tested on different types of objects, systems and countermeasures. For the experiment a dataset (~ 10 million) was used, containing information about objects with various threats. The experimental data showed the choice of optimal countermeasures distributed as follows: 44.37% - antivirus filtering tools, 31.72% - filtering in the browser, 23.04% - notification to the resource administrator; 0.86% - filtering by the parental control system. The experiment demonstrated the possibility of their implementation and applicability of the developed technique for the selection of countermeasures. The technique can be applied in public control systems (schools, educational institutions), in the corporate segment (assistance in solving the restriction of access to information resources in the company's network), as well as for personal use (parental control systems, filtering).

References
[1] Desnitsky V A, Kotenko I V and Parashchuk I B 2019, May methods of assessing the effectiveness of network content processing systems for detecting malicious information taking into account the elimination of uncertainty in the semantic content of information objects 2019 SCM. (St. Petersburg, Russia), pp 41-44
[2] Kotenko I V and Saenko I B 2012 Developing the system of intelligent services to protect information in cyber warfare (SPIIRAS Proceedings vol 3(22)), pp 84-100.
[3] Shiang L 2015 Systems and methods for media detection and filtering using a parental control logging application. U.S. Patent 9,204,193, issued December 1, 2015
[4] Schepis A and Newstadt K 2018 Methods and systems for applying parental-control policies to media files. U.S. Patent No. 9,330,274. Washington, DC: USPTO, 2016
[5] Slemmer J B, Olsson S, Kreiner B, Jarboe A and Kleinfelter K 2018 Parental control in a networked environment. U.S. Patent Application 16,115,624, filed December 27, 2018
[6] Schepis A and Newstadt K 2016 Systems and methods for applying parental-control approval decisions to user-generated content. U.S. Patent 9,495,547, issued November 15, 2016
[7] Vinayakumar R, Soman K P and Poornachandran P 2018 Evaluating deep learning approaches to characterize and classify malicious URL’s., Journal of Intelligent & Fuzzy Systems, 3 pp 34 1333-1343
[8] Agbefu R E, Hori Y and Sakurai K 2013 Domain information based blacklisting method for the detection of malicious webpages. IJCSDF, 2(2), pp 36-48
[9] Mehta R P, Sanghvi M A, Shah D K and Singh A 2020 Sentiment analysis of tweets using supervised learning algorithms. In ICTSCI proceedings, Springer, Singapore, pp 323-338
[10] Sarma P K and Sethares W 2018 Simple algorithms for sentiment analysis on sentiment rich, data poor domains, In Proceedings of the COLING, pp 3424-3435
[11] Schouten K, Frasincar F, Dekker R and Riezebos M 2019 Heracles: a framework for developing and evaluating text mining algorithms, Expert Systems with Applications, 127, pp 68-84
[12] Larkin A D and Tushkanov V N 2018 A way to limit user access to suspicious objects of a social network. Patent RU No. 2651252. Patent holders: Kaspersky Lab
[13] Fabrikant A and Sarma A D 2014 Pre-processing of social network structures for fast discovery of cohesive groups. U.S. Patent No. 8,909,646. Washington, DC: USPTO

Acknowledgments
The work is performed by the grant of RSF #18-11-00302 in SPIIRAS.