Web Page Classifier Information System as an information software system reducing Internet access by usage of keywords related with 'weight'

Piotr Wójcik
Polish-Japanese Academy,
Warsaw,
pwojcik@pja.edu.pl

August 8, 2018

1 Abstract

This article describes a proposal of a new method to scan Internet resources for pages that are unsafe for selected Internet users. Currently, it is a plan for a prototype that will be evaluated for performance and usability. The method uses the concept of weighted keywords to classify and then even censor web pages. The system will block selected Internet web pages by updating Domain Names System.

Keywords: *Internet, classifier, classification, keywords with wage, Domain Names System, DNS*

2 Introduction

Almost every indexed web page can be classified by usage of particular**keywords**. Analyzing appearance of keywords used on page known also as **search phrase** can tell about age requirements of web surfer for selected web page. Some phrases sets can tell the rule engine directly, others can only guide for proper category. For example if on page appear word "Nazism" there is a high probability that page is not appropriate for the fourteen year old's or youngers. The classification of this type can be described as **direct rule**. Other phrases only guide us for adequate classification. Let's think about phrase "Adolf Hitler". There is still huge probability that web page contains text about Nazism, but also can lead to page about Adolf Hitler's biography that can be suitable for high school students.

This leave us with important problem, dysfunction of proper phrases classification that appears on web pages. Let's think about word "breast", which could lead us to erotic or even pornographic pages. But when we add another keyword and find material about "breast cancer" we can find medical pages about common disease that can be discussed also by high school students. However this time adding another keyword can lead us to very different pages in comparison to phrase "Adolf Hitler". We would call it **not direct rule**. This forces us to careful selection of keywords like proper phrase weight to distinguish direct form of non direct rules of comparison.

Classification could be done by artificial neural networks which can distinguish pages with nudity on images or videos [1–3]. The technique of image extraction and accurate skin detection from web pages is very popular and gives good results in Internet media classification [4,5] but in this article

I won't focus on that topic.

On the other hand there is unpopular concept connected to this material and is called Internet Censorship. It was discussed in Jonathan Zittrain, Benjamin Edelman [6] none regime countries like Australia, Great Britain and United States of America are constantly using this controversial technique. The main reason of this idea is to choose lesser of two evils which principle comes down to safety of young Internet surfers [7, 8].

3 DNS

The best as well as the most safe way to limit access to web pages is basicly to prepare appropriate Domain Name System - DNS. [9–18]. Technique discussed in this article is called **the Exclusion Filtering** [5].

Of course there is an easy way to bypass all this security features and simply not search for domain name IP address by directly typing this number. This leads us to necessity of remembering that addresses. In IP protocol version four we encounter to 256^4 different numbers which is huge number of potential addresses but in upcoming IP protocol version six we could address unthinkable number $3.4*10^{1038}$. In practice those big numbers we should reduce by number of sub networks of address classes that not lead directly to web page.

Another way of bypass this security feature is to manually replacing DNS server address in the protected computer but even this step requires administrative privileges.

Alg 1. The basic Internet web page classifier algorithm

```

//input:
//     phrase - analysed phrase,
//     analysedPage - web page with URL and other characteristics,
//     compromisedWebPages - collection of compromised web pages,
//     redFlagPhraseDictionary - collection of phrases
//                                 that compromise web page
//output:
//     score - points of page classification
//     compromisedWebPages - modified input object

mostCompromisedPages = compromisedWebPages.top();
foreach (phrase in webPageTextContent)
{
  if((phrase in redFlagPhraseDictionary)
      || (phrase like redFlagPhraseDictionary))
  {
    redFlagPhraseDictionary.computeWage(phrase, mostCompromisedPages);
    score += redFlagPhraseDictionary.wage(phrase);
    compromisedWebPages += analysedPage;
  }
}
```

4 Keywords with weight

Concept of **keywords with weight** is a simple idea of extending semantic meaning of the word that is written in text, For instance usage of Latin alphabet letters by number value. To add more complications let analyze for example word sex which may function as a verb or function as a noun with five different meaning according to Dictionary.com [19]. This complicate direct classification because having one world we cannot decide with correct probability if the page is safe so we might want to estimate **dangerous factor** ω as equal 0.6. Lets define **keyword with weight** K as:

$$K = \{a_1a_2a_3...a_n, \omega : a_n \in [A - Z],$$

$$n \in \mathbb{N}, \omega \in <0,1>\}.$$

The bigger is dangerous factorω the more valuable keyword is because is it more efficient in classifying page as unsafe. In Red Flag Dictionaries described in algorithm [Alg1] we are expecting to include those keywords.

5 Keywords weight update and searching for new valuable keywords

Main characteristics of a language is that, that is dynamically changing. More new words appearing or slogan words are used. Some of them become obsolete which rises issue to constantly update a dictionary and weight of dangerous factors ω.

Other fact is that similar pages use similar phrases in content. Traversing base set of pages can provides as not only an updated of weights but also can give us new set of phrases. In search of new phrases is applied more complex process than updating dangerous factors in existing set of keywords.

For the brief understanding of dynamic, update of dangerous factors ω parameters and classification of Internet pages I present this simple algorithm [Alg1] for future analysis and implementation. Above code is implemented in C# [20] like language. The only not compatible feature is **like** keyword that is not implemented in that language. This function is implemented in 4GL languages like SQL and can be easily replaced by calling method *like()*. The body of this function may for example contain *Soundex* algorithm created by Robert Russell and Margaret Odell [21]. The result of this algorithm is four-digit code containing information about phonetic similarity of two given on input words.

Idea of this algorithm depends on analysis of all *phrases* on given Internet page *analysedPage* to find any similarities with keywords in set *redFlagPhraseDictionary* (line 15) called **red flag phrases**. If phrase is similar or identical then:

1. update all dangerous factor parameters in *RedFlagPhraseDictionary* classification set and update the ranking static class *MostCompromisedPages* (line 18),
2. sum up the *score* for this page (line 19),
3. attach this page to list of compromised pages (line 20).

On the line 12 we are setting up the top list of the most compromised pages *compromisedWebPages.top()* as a source included in every iteration for this algorithm.

6 Web pages category

We may propose grouping of an Internet web pages in 12 different categories:

- information pages,
- web databases,
- commercial pages (shops, e-business etc.),
- social networks,
- religious pages,
- dating sites,
- erotic pages,
- pornographic pages,
- risk sites (for example online casinos),
- sites with hate speech,
- sites with illegal materials and
- sites with hate and violence.

This might not be the final list of World Wide Web categorisation and extension may required. Because of the context of pages the best current strategy is to deal with classification how to use different dictionaries, parameters or even algorithms to more efficient score decision. In case of classified site as unsafe, proper description should be added to database for future analysis and performance issues.

7 Web Page Classification Information System modules

The proposed Web Page Classification Information System WPCIS is going to be built with several databases and required several different algorithms to process information.

First module contains database with all discovered Internet Pages that going to be process, was visited by one or many different classification algorithm, classified page or fully classified page with all external link processed. In case of classified or fully classified page description will be attached containing information which keywords was used to compute score of this page, the value of dangerous factors parameters used to make such index and phrases enumeration on this pages.

The Second module is going to be build with much smaller database containing perspectives of actual and used in the past sets of dictionaries and phrases with weight. Each perspective are going to have date of introduction, version number and dictionary values and may be withdraw at any time in case of new parameters introduction. This database is a heart of system and going to be heavily updated by processing program.

Third, the smallest but still important module is a program containing the algorithm to compute dangerous factors ω parameters for red flag phrases.

Fourth important module is a creator and updater of DNS server which use base of first module data. In some cases classification of pages could change so it is important to give this change possibility for DNS records.

Fifth the most complicated and resources consumer module is data-mining program that extracts new phrases for red flag dictionary algorithms.

Of course this system is going to grow in time and some modules are going to be divided with smaller components with more specialized features like fourth module which is going to be divided with base for different DNS servers grouped by different pages categories described in section Web pages category.

8 Web Page Classification Information System appliance

First and the most obvious place to use this appliance is home environment where young children and adolescent require safe Internet access. Second also important place to use the system is in work environment and educational places like schools, museums where it is important to provide reduced access to some Internet resources.

9 Conclusion and future plans

The reason of this article appearance is to design new classification method. In near future new systems are going to be implemented. The most important feature of this program is to create plenty of DNS servers containing records of censured web domains depending on the outcome of various classification algorithms. This DNS systems can be applied to endpoint of computers and are going to greatly reduce access to unsafe or prohibited web resources.

10 References

[1] Will Archer Arentz , Bjørn Olstad. Classifying offensive sites based on image content.

[2] Radhouane Guermazi , Mohamed Hammami, Abdelmajid Ben

Hamadou. Combining classifiers for web violent content detection and filtering.

[3] Giuseppe Amato , Pablo Bolettieri , Gabriele Costa , Francesco la Torre , Fabio Martinelli. Detection of images with adult content for parental control on mobile devices.

[4] Mohammad Reza Mahmoodi. High performance novel skin segmentation algorithm for images with complex background.

[5] Paul Greenfield, Peter Rickwood, Huu Cuong Tran. NetAlert and the australian broadcasting authority.

[6] Jonathan Zittrain, Benjamin Edelman. Internet filtering in china.

[7] Piotr Łuczuk. *Cyberwojna*.

[8] John G. Palfrey, Jr. Four phases of internet regulation.

[9] J. Postel, J. Reynolds. Domain requirements.

[10] P. Mockapetris. Domain Names - Concepts and Facilities.

[11] P. Mockapetris. Domain names - implementation and specification.

[12] Yakov Rekhter, Susan Thomson, Jim Bound, Paul Vixie. Dynamic updates in the domain name system (DNS UPDATE).

[13] R. Elz, R. Bush. Clarifications to the DNS specification.

[14] D. Eastlake, 3rd, C. Kaufman. Domain name system security extensions.

[15] Donald E. Eastlake 3rd. DNS request and transaction signatures (SIG(0)s).

[16] Brian Wellington. Secure domain name system (DNS) dynamic update.

[17] Brian Wellington. Domain name system security (DNSSEC) signing authority.

[18] Edward Lewis. DNS security extension clarification on zone status.

[19] Dictionary.com.

[20] Andrew Troelsen. *Jezyk C# 2010 i platforma .NET 4*. Apress.

[21] Donald E. Knuth. *The Art of Computer Programming*, wolumen 3.