Web Page Classification Information System as an information software system reducing access to Internet resources using keywords with wages

Piotr Wojcik
Polish-Japanese Academy,
Warsaw,
piotr.wojcik@pja.edu.pl

Krzysztof Stencel
Warsaw University,
Warsaw,
stecel@mimuw.edu.pl

July 22, 2018

1 Abstract

Article describes proposition of new method of scanning Internet resources to find unsafe web pages for Internet users. Actual it is only an plan for the prototype that needs to be measured for his performance and usability. It is using concept of keywords with wages to classify or even selected censorship access to Internet thanks to list of blocked Internet web pages on dynamic updating Domain Names Servers.

Keywords: *Internet, classificator, classification, keywords with wage, Domain Names Server, DNS*

2 Introduction

Almost every indexed Internet web page can by classified using proper **keywords**. Analysing appearance of keywords used on page known also as **search phrase** can tell about age requirements of Internet surfer for selected web page. Some phrases can tell the rule engine directly others can only guide for proper category. For example if on page exist word "nazism" there is a high probability that page is not appropriate for the fourteen years old or younger pearson. The classification of this type can be described as **direct rule**. Other phrases only guide us for proper 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 Hitlers biography that can by suitable for high school students.

That face us to the important problem 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 named it **not direct rule**. This forces us to careful keywords grouping and establishing proper phrase wages to distinguish direct from 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]. This technique of finding percent of coverage of skin color is very popular and gives good results in Internet media classification [4]. However this topic will not be discussed in this article.

There is other unpopular concept connected to this material and it is called Internet censorship. It was discussed in Jonathan Zittrain, Benjamin Edelman article [5]. However more and more none regime countries like Australia, Great Britain and United States of America are using this controversial technique. The main reason of this idea is to choose lesser of two evils principle when there comes to talk about safety of younger Internet users [6].

3 DNS

The best and also the most safe form to limit access to discover world wide web pages is to prepare appropriate Domain Names Servers DNS [7–12].

Of course there is a easy way to bypass this security feature and not search for domain name IP address by directly typing this number. This leads us to necessity to remember this addresses. In IP protocol version four we encounter to 256^4 different numbers with it 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 numbers we should reduce by number of sub networks of addresses classes that not lead directly to web pages content.

Another way of bypass this security feature is to manual replacing DNS server address in protecting computer but even this step requires administration privileges.

Alg 1. Podstawowy algorytm klasyfikacji stron internetowych

```

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

Under the concept of **keywords with wage** are simple idea of extending semantic meaning of the word that are written for example with use of Latin alphabet letters by number value. For example word ”sex” that may function as a verb it function also as a noun with five different meaning according to Dictionary.com [13]. This complicates a little direct classification because having this one world we cannot decide with great probability if the page is safe so we might want to estimate it **dangerous factor** ω as equal 0.6. Let's define **keyword with wage** K as:

$$K = \{a_1a_2a_3...a_n, \omega : a_n \in [A - Z],$$
$$n \in \mathbb{N}, \omega \in <0, 1>\}.$$

The larger is dangerous factor ω the more valuable is keyword because it is more efficient in classifying page as unsafe. In Red Flag Dictionaries described in algorithm [alg1] we are expecting to include those keyword.

5 Keywords wages update and searching for new valuable keywords

One of the characteristic of language is that he is dynamically changing and changing over the years. More and more words are appearing. Some of them are become obsolete. This rises issue to constantly update a dictionary and wages of dangerous factor. Other fact is that similar pages use similar phrases in content. Traversing base set of pages can provides as not only an updated of wages but also can give us new set of phrases. In search of new phrases is applied more complex process than updating dangerous factor in existing set of keywords.

For the brief understanding of dynamic update dangerous factor parameter and classification of Internet pages I present simple algorithm [Alg1] for future analysis and implementation. Above code is implemented in C# like language. The only not compatible feature is an **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 method may for example contain Soundex algorithm created by Robert Russell and Margaret Odell [14]. The result of this algorithm is four-digit code containing information about phonetic similarity of two 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,
- 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 be not a final list o World Wide Web categorization and maybe requires an extension. Because of context of this pages the best strategy to deal with classification is to use different dictionaries, parameters or even algorithms to more efficient score decision. In case of classify site as a unsafe proper description should be added to database for future analysis and performance issues.

7 Web Page Classification Information System modules

8 Web Page Classification Information System appliance

9 Conclusion and future plans

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

[6] Piotr Łuczuk. *Cyberwojna.*

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

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

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

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

[11] R. Elz, R. Bush. *Clarifications to the DNS Specification.*

[12] D. Eastlake. Domain name system security extensions.

[13] Dictionary.com.

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