Improving the Categorization of Web Sites by Analysis of Html-Tags Statistics to Block Inappropriate Content

Dmitry Novozhilov, Igor Kotenko and Andrey Chechulin

Abstract The paper considers the problem of improving the quality of web sites categorization using data mining methods. This goal is important for automated systems of parental control. The purpose of such systems is protection from unwanted or inappropriate information. The novelty of the proposed approach is in usage of HTML tags statistics of web pages to improve the categorization of sites that are similar in terms of textual content, but differing in their structural features. The paper describes the architecture of the categorization system, the algorithm of its work, the results of experiments, and assessment of classification quality.

1 Introduction

Data mining methods are known to be used for detecting hidden knowledge in the data: unknown, non-trivial and practically useful. During data analysis it is often necessary to classify the studied object to one of predefined classes; this is the classification problem. Its correct solution is very important and leads to significant progress in many areas. The distinctive features of our time are continuous development and ubiquity of the Internet. In such conditions the importance of automatic classification systems, that distribute web pages by category and block those that are undesirable or offensive, increases. This is extremely important, for example, to protect children from sites containing inappropriate content or to counteract the spread of malware and pirate content.

D. Novozhilov · I. Kotenko (✉) · A. Chechulin
Laboratory of Computer Security Problems, St. Petersburg Institute for Informatics
and Automation (SPIIRAS), 39 14 Linija, St. Petersburg, Russia
e-mail: ivkote@comsec.spb.ru

D. Novozhilov
e-mail: novozhilov@comsec.spb.ru

A. Chechulin
e-mail: chechulin@comsec.spb.ru

P. Novais et al. (eds.), *Intelligent Distributed Computing IX*,
Studies in Computational Intelligence 616,
DOI 10.1007/978-3-319-25017-5_24

There are many different approaches to the web sites classification. The most efficient and widely used is the analysis of the text content of web pages. However, there are some categories of sites (forum, blog, news) which are almost the same in text content, whereas their structural features are different. In such situations researchers use other classification methods. For example, you can parse URL addresses of the pages or HTML tags of their markup. One of the options in the last approach is to check the presence/absence of specified tags. In the paper we propose an original approach, which, unlike the existing ones, is based on the analysis of statistics of HTML tags, which is the ratio of all occurrences of the tag to the total number of tags on the site. The main objective of the paper is to improve the quality of classification for web sites, which classification by text is difficult, by analysis of web sites based on information about their HTML tags. All stages of this research are reflected in this paper, which has the following structure: at first the review and analysis of related work is conducted, then the proposed solution, the experiments and their results are described. The last section is devoted to conclusions and future plans.

2 Related Works

The most widely applied method is the classification according to the text [1–3]. Another alternative is to move from a consideration of the documents as sets of words to the analysis of their meanings, which are taken from the lexical database [4]. Disadvantage of text classification is that it does not take into account web pages particularities: HTML document is linked by references to other documents; it contains images and other non-text elements. Difficulties are also caused by categories with similar text content, but differing in their structure (for example, blogs, forums, chats). Thus, the method based on the URL analysis was developed. Using it we assume that the web page is rarely visited unless it generates interest among potential readers. That means that the address of the site should somehow reflect its theme [5]. One of the methods here is to split the URL into its component parts, and to analyze the parts obtained. This approach is implemented in [6] for URL analysis to protect from phishing sites. The position of a particular portion in the site address is also important. Another way is to use the length of the host name and the entire URL, count the number of different symbols and analyze URL fragments between these characters. Besides that there are also signs on the basis of information about the host (geographical features, date of registration, the value of TTL, etc.). All these attributes are fed to the input of any classifier [7]. In [8] and [9] there are references to the method associated with the sequence analysis of *n*-grams, for which the frequency of appearance is calculated. To identify categories, based on structural characteristics, it is necessary to look for other methods, one of which is the usage of HTML tags. The information contained in tags like <title> or <meta> together with the text content may serve as an important source for analysis [10–13]. Our paper discusses a new approach that is based on the analysis not of the content or the number of HTML tags on the page, but their statistics, which is defined as the ratio of all occurrences of the tag to the total number of tags on the site.

3 Models and Implementation

Web pages are known to be different from regular documents primarily by the fact that they are semi-structured using HTML markup tags, interconnected with links, contain fragments of code that is executable on the server side and on the client. Therefore, it is to use methods, which take into account the particularity of the analyzed data. One of possible solutions may be the approaches based on structural features of web pages, i.e. with HTML tags.

The proposed method is also not based on the stored source code of the web pages for later analysis, but works with the statistics of HTML tags instead. We understand the statistics S of tags as the totality of their occurrence frequencies f_{i}, which is defined as the ratio of the number of instances of this tag n to the total number N of tags on the page, expressed as a percentage. The result is rounded to be more informative: $S = \cup f_i; f_i = \frac{n_i}{N} * 100\,\%$. We should note that such solution was not found in the beginning of our research, and at first we analyzed simply the number of tags of each type on the page. However, this approach is not quite correct, because, for example, it is incorrect to compare 100 tags <div> on the pages, consisting of 250 and 1000 tags, and they point to a completely different result. The final classifier is based on the Naïve Bayes and Decision Tree algorithms, which basic predictions are combined on the upper level using stacking. As the base models, stacking uses various classification algorithms trained on the same data. Then the meta-classifier is trained on the input data, supplemented by the results of the forecast of the base algorithms. The idea of stacking is that the meta-algorithm learns to distinguish which of the base algorithms it should "trust" on which areas of input data.

To evaluate the quality of classification we use such metrics as precision, recall, accuracy, and the F-measure which combines information on the precision and recall. It should be noted that for the class of systems that perform parental control accuracy metric is of particular importance, since a large number of false positives may cause refusal from application of such systems.

The task of finding the frequency of tags can be solved in several different ways. Let us select two of them: (1) search the tags in the entire HTML document and count the number of occurrences of each of them; (2) use representation of the HTML document as a tree that significantly simplifies solution of the problem, providing a variety of navigation functions and access to its elements. One of the arguments to choose the second approach was that such a tree data structure already exists—it was built for the needs of document parsing and storing its text content to a file without HTML tags.

The key is the tag name, and values are all its representatives, the number of which is counted. If we apply the function to the root of the tree, we shall obtain problem solution. For further analysis, all tags are used which frequency exceeds 2 %. It is empirically set value that allows to exclude from consideration tags that are common to all pages, such as <HTML> , <title> , <head> , <body> etc.

4 Experiments

The experiments were conducted on two data sets ("set1" and "set2"). "Set1" was created basing on the original data from URLBlacklist.com [14], which included the categories: "Books, Hunting, News, Dating, Guns". For each category 1 thousand sites were selected. "Set2" includes content of URLBlacklist.com [14], combined with part of the categories taken from the list of "Shalla Secure Services KG" [15]. Several different catalogs are used to find common features of the sites, as well as by the lack of source data for certain categories within one source and sufficient number of them in another one. Finally there were selected 13 categories: "Chat, Drugs, Forum, Guns, Hunting, Jobsearch, Magazines, Medical, Movies, Music, Press". In each of them there were about 1.5 thousands of sites.

In the course of source data preparation the attention was paid to the boundaries of the categories. Such heterogeneous categories as "radio-tv" or "audio-video" were excluded from consideration, because in fact each of them was divided into two others. "Drugs" and "medical", "guns" and "hunting", on the contrary, were taken specifically to see the work in cases where some specific words and combinations can be common for them. The experimental results for "set1" are presented in Fig. 1. Classification results for tagged and textual features for the first set are shown in Fig. 2. The results of the experiments for the second set are presented in Fig. 3. Classification results for tagged and textual signs for "set2" are shown in Fig. 4. Analysis of the experimental results allows to make conclusion about generally low quality, which does not allow to use this method as the primary tool. The "set 1" gives higher value of accuracy equal to 35.43 %, because it contains less "controversial categories" which may intersect (only "guns" and "hunting"). For the "set 2" the accuracy decreases to 15.08 %, while it is clearly visible that categories "press" and "hunting" became "absorbing".

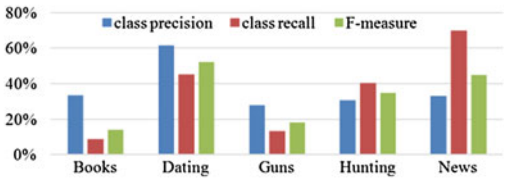

Fig. 1 The values of basic metrics for the set "Set 1"

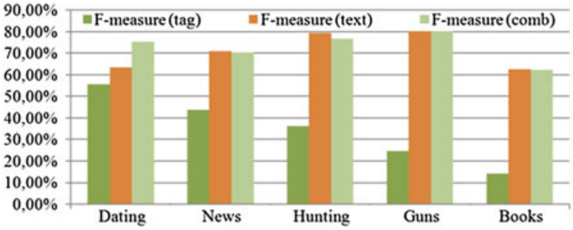

Fig. 2 Categories for set "Set 1", ordered by descending F-measure for classification by tags

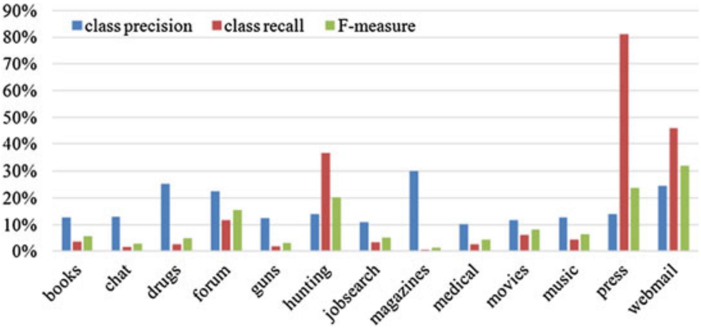

Fig. 3 The values of basic metrics for the set "Set 2"

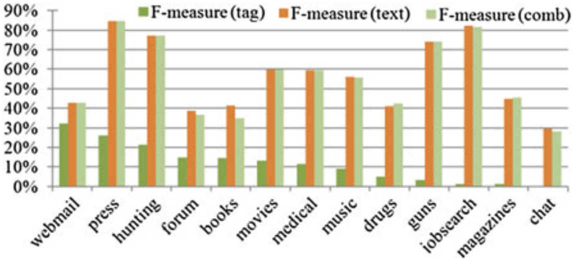

Fig. 4 Categories of set "Set 2", ordered by descending F-measure for classification by tags

Figure 5 shows the values of accuracy for approaches based on the analysis of tags, text and their combinations. The results reflect the improvement of the quality of classification by combining these approaches for 5 and 13 categories. Thus, the studies show that the proposed approach based on the statistics of the HTML tags does not solve the problem of rating by itself, but it can be a good addition in the outlining of categories that differ in their structural features.

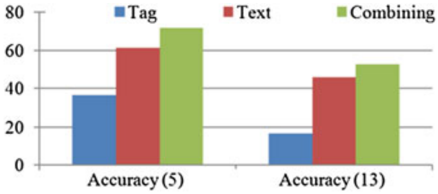

Fig. 5 The accuracy for approaches based on different features and their combination

5 Conclusion

This paper discussed approaches to categorizing web pages that do not have significant differences in text classification, but having different structure. The essence of the proposed method is the use of HTML tags statistics, which is fed to the classifiers. The obtained results show that the quality of the tags based classification is not sufficient to apply this method as a standalone one. But it can be used as a useful complement to existing systems with textual classification. The principles investigated can be applied to improve the quality of systems for protection from information, such as the parental control systems. Further research directions are connected with search for other classifiers and their combinations that will allow to combine textual and statistics tags data analyses, to get rid of "absorbing categories" that are characteristic to decision trees.

Acknowledgment This research is being supported by The Ministry of Education and Science of The Russian Federation (contract # 14.604.21.0147, unique contract identifier RFMEFI60414X0147).

References

1. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: ECML-98, LNCS, vol. 1398, pp. 137–142. Springer (1998)
2. Ko, Y., Seo, J.: Automatic text categorization by unsupervised learning. In: Coling'00, pp. 453–459. Morgan Kaufmann (2000)
3. Ntoulas, A., Najork, M., Manasse, M., Fetterly, D.: Detecting spam web pages through content analysis. In: ACM, pp. 83–92 (2006)
4. Kehagias, A., Petridis, V., Kaburlasos, V.G., Fragkou, P.: A comparison of word-and sense-based text categorization using several classification algorithms. J. Intell. Inf. Syst. **21** (3), 227–247 (2000)
5. Attardi, G., Gulli, A., Sebastiani, F.: Automatic web page categorization by link and context analysis. In: THAI'99, pp. 105–119 (1999)
6. Khonji, M., Iraqi, Y., Jones, A.: Enhancing phishing E-Mail classifiers: a lexical URL analysis approach. Int. J. Inf. Secur. Res. **6**, 236–245 (2012)
7. Ma, J., Saul, L.K., Savage, S., Voelker, G.M.: Beyond blacklists: learning to detect malicious web sites from suspicious URLs. In: KDD'09, pp. 1245–1254. ACM (2009)

8. Kan, M.-Y., Thi, H.O.N.: Fast webpage classification using URL features. In: ICIKM 2005, ACM (2005)
9. Geide, M.: N-gram Character Sequence Analysis of Benign vs. Malicious Domains/URLs. Available at http://analysis-manifold.com/ Accessed 24 March 2015
10. Meshkizadeh, S., Masoud-Rahmani, A.: Webpage classification based on compound of using html features and url features and features of sibling pages. Int. J. Adv. Comput. Technol. **2**(4), 36–46 (2010)
11. Patil, A.S., Pawar, B.V.: Automated classification of web sites using naive bayesian algorithm. In: IMECS2012, vol. 1, p. 466 (2012)
12. Riboni, D. Feature selection for web page classification. In: EURASIA-ICT-2002 (2002)
13. Kotenko, I., Chechulin, A., Shorov, A., Komashinsky, D.: Analysis and evaluation of web pages classification techniques for inappropriate content blocking. LNAI **8557**, 39–54 (2014)
14. URLBlacklist.com.: http://urlblacklist.com/ Accessed 24 March 2015
15. Shalla Secure Services KG.: http://www.shallalist.de/ Accessed 24 March 2015