Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

Stable web spam detection using features based on lexical items

Marcin Luckner [a],*, Michał Gad [b], Paweł Sobkowiak [c]

[a] Faculty of Mathematics and Information Science, Warsaw University of Technology, Koszykowa 75, 00-662 Warszawa, Poland
[b] EO Networks, ul. Głuszycka 5, Warszawa, Poland
[c] Sensi Soft, ul. Głuszycka 5, Warszawa, Poland

ARTICLE INFO

Article history:
Received 2 February 2014
Received in revised form
29 June 2014
Accepted 16 July 2014
Available online 24 July 2014

Keywords:
Web spam detection
Spam detection features
Regular expressions
Context analysis
Lexical items analysis

ABSTRACT

Web spam is a method of manipulating search engines results by improving ranks of spam pages. It takes various forms and lacks a consistent definition. Web spam detectors use machine learning techniques to detect spam. However, the detectors are mostly verified on data sets coming from the same year as the learning sets. In this paper we compared Support Vector Machine classifiers trained and tested on WEBSPAM–UK data sets from different years. To obtain stable results we proposed new lexical-based features. The HTML document – transformed into a text without HTML tags, a set of visible symbols, and a list of links including the ones from tags – gave information about weird combinations of letters; consonant clusters; statistics on syllables, words, and sentences; and the Gunning Fog Index. Using data collected in 2006 as a learning set, we obtained very stable accuracy among years. This choice of the training set reduced the sensitivity in 2007, but that can be improved by managing the acceptance threshold. Finally, we proved that the balance between the sensitivity and the specificity measured by the Area Under the Curve (AUC) is improved by our selection of features.

1. Introduction

The detection of spam emails was – and still is – a serious problem for the Internet community (Carpinter and Hunt, 2006). The new spam filtering approaches have raised the hope of reducing this problem (Filasiak et al., 2014), but the detection of new types of spam – SMS spam (Xu et al., 2012), MMS spam (Yoon et al., 2010), video spam, and web spam (Potdar et al., 2010) – is still one of the most challenging issues. The most common type in the list is web spam that exploits vulnerabilities and gaps in the web 2.0 to inject links to spam content into dynamic and sharable content such as blogs, comments, reviews, or wikipages.

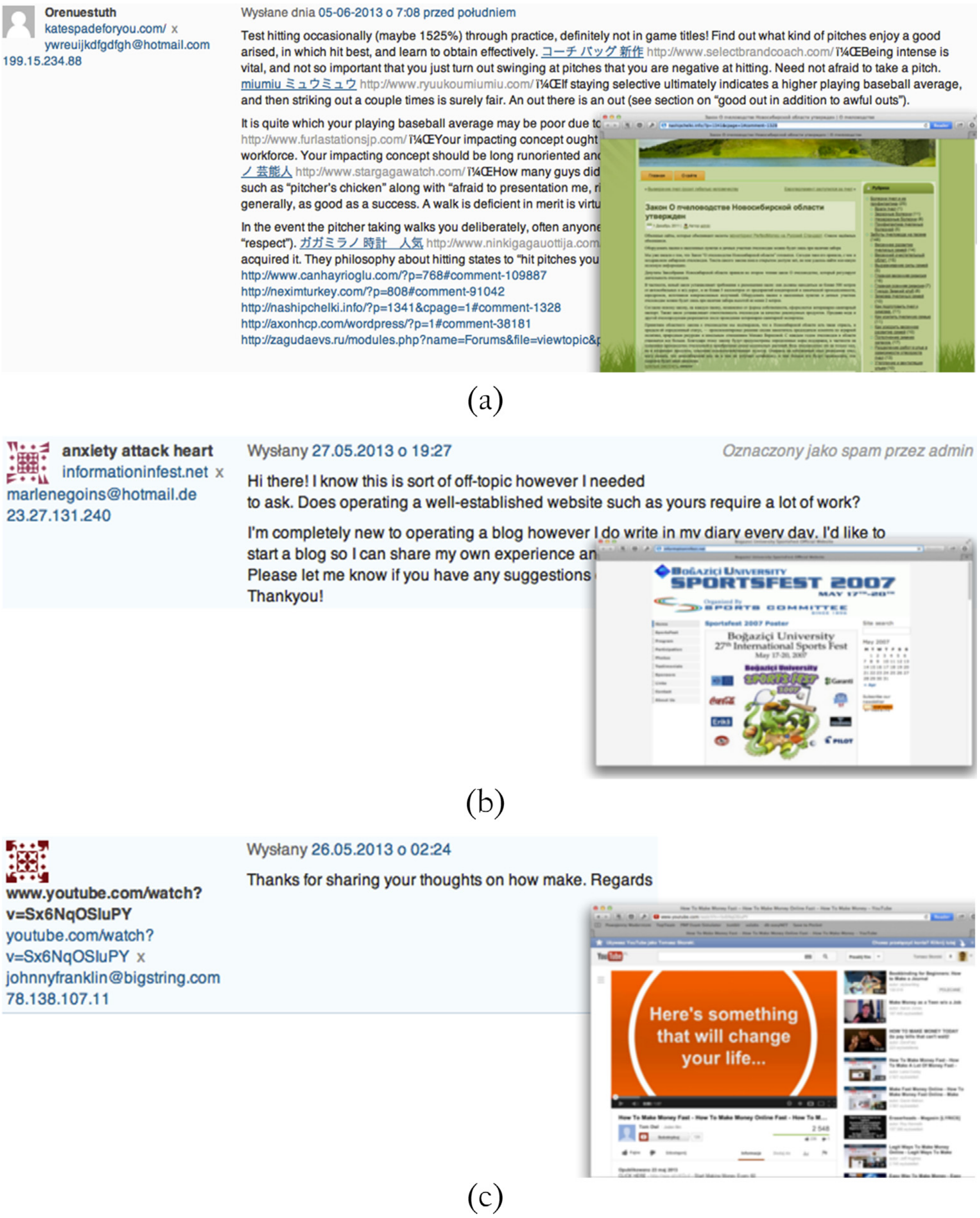
Web spam takes various forms and lacks a consistent definition. Therefore, web spam detectors use machine learning techniques to create a model of spam from training sets that include spam and non-spam examples. WEBSPAM–UK collections from 2006 to 2007 are very popular training sets (Castillo et al., 2006). Many projects have used these data sets to test web spam detectors. However, most detectors were verified and proved effective only on data coming from a single year (c.f. Mahmoudi et al., 2010; Algur and Pendari, 2012). One may conclude that the presented solutions are temporary.

* Corresponding author.
E-mail addresses: mluckner@mini.pw.edu.pl (M. Luckner), michal.gad@eo.pl (M. Gad), psobkowiak@acm.org (P. Sobkowiak).
http://dx.doi.org/10.1016/j.cose.2014.07.006
0167-4048/© 2014 Elsevier Ltd. All rights reserved.

Fig. 1 shows several web spam injections into web forum comments. Fig. 1(a) presents a typical example of web spam comment with several injected links. This spam was created to promote link farms and provide credibility to the spammer website. The second example of text spam on a web blogging platform Fig. 1(c) refers to the image. Any human can recognise the linked page as spam, but an automaton needs special techniques of image spam detection (Gao et al., 2009; Wakade et al., 2013). The link contained in the last example of pingback spam on Wordpress platform Fig. 1(b) directs users to the Youtube spam movie. Because of computation costs, video spam is very hard to detect and approaches to detect video spam are still being developed (Luz et al., 2012).

The aim of this study was to create a web spam detector that works over years. We selected several features based on lexical items and distinguished web spam from non-spam context. We believe that such features could be common for the WEBSPAM–UK2006 and WEBSPAM–UK2007 data sets. To test this hypothesis we created three web spam detectors: trained and tested using 2006 data; trained and tested using 2007 data; trained using 2006 data, but tested using 2007 data. The results of the first and the second detector formed points of reference for the third one. We expected that the accuracy obtained by the last detector would be similar to the accuracy obtained using 2006 data. The results obtained using 2007 data allowed us to estimate the reduction of the accuracy obtained by the third detector.

To estimate the influence of the selected features, we repeated the test using only commonly used features. Our hypothesis was that the obtained results would be worse than the ones achieved in the first test.



Orenuestuth
katespadeforyou.com/ x
ywreuijkdfgdfgh@hotmail.com
199.15.234.88

Wysłane dnia 05-06-2013 o 7:08 przed południem

Test hitting occasionally (maybe 1525%) through practice, definitely not in game titles! Find out what kind of pitches enjoy a good arised, in which hit best, and learn to obtain effectively. コーチ バッグ 新作 http://www.selectbrandcoach.com/ ï¼ŒBeing intense is vital, and not so important that you just turn out swinging at pitches that you are negative at hitting. Need not afraid to take a pitch. miumiu ミュウミュウ http://www.ryuukoumiumiu.com/ ï¼ŒIf staying selective ultimately indicates a higher playing baseball average, and then striking out a couple times is surely fair. An out there is an out (see section on "good out in addition to awful outs").

It is quite which your playing baseball average may be poor due to
http://www.furlastationsjp.com/ ï¼ŒYour impacting concept ought
workforce. Your impacting concept should be long runoriented and
ノ 芸能人 http://www.stargagawatch.com/ ï¼ŒHow many guys did
such as "pitcher's chicken" along with "afraid to presentation me, ri
generally, as good as a success. A walk is deficient in merit is virtu

In the event the pitcher taking walks you deliberately, often anyone
"respect"). ガガミラノ 時計　人気 http://www.ninkigagauottija.com
acquired it. They philosophy about hitting states to "hit pitches you
http://www.canhayrioglu.com/?p=768#comment-109887
http://neximturkey.com/?p=808#comment-91042
http://nashipchelki.info/?p=1341&cpage=1#comment-1328
http://axonhcp.com/wordpress/?p=1#comment-38181
http://zagudaevs.ru/modules.php?name=Forums&file=viewtopic&p

(a)

anxiety attack heart
informationinfest.net x
marlenegoins@hotmail.de
23.27.131.240

Wysłany 27.05.2013 o 19:27

Oznaczony jako spam przez admin

Hi there! I know this is sort of off-topic however I needed
to ask. Does operating a well-established website such as yours require a lot of work?

I'm completely new to operating a blog however I do write in my diary every day. I'd like to
start a blog so I can share my own experience an
Please let me know if you have any suggestions
Thankyou!

(b)

www.youtube.com/watch?
v=Sx6NqOSluPY
youtube.com/watch?
v=Sx6NqOSluPY x
johnnyfranklin@bigstring.com
78.138.107.11

Wysłany 26.05.2013 o 02:24

Thanks for sharing your thoughts on how make. Regards

(c)

Fig. 1 – The examples of web spam comments with linked content. The links under signatures lead to the web sites presented in the bottom right corners.