Expert Systems with Applications xxx (2015) xxx–xxx

Contents lists available at ScienceDirect

Expert Systems with Applications

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

A dynamic model for integrating simple web spam classification techniques

Jorge Fdez-Glez [a], David Ruano-Ordas [a], José Ramón Méndez [a,b], Florentino Fdez-Riverola [a,b], Rosalía Laza [a,b], Reyes Pavón [a,b,*]

[a] Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, University of Vigo, 32004 Ourense, Spain
[b] CITI: Centro de Investigación, Transferencia e Innovación, Parque Tecnológico de Galicia – Tecnópole, Avda. de Galicia N° 2, San Ciprián das Viñas, 32900 Ourense, Spain

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Web spam
Dynamic model
Instance based reasoning
Ensemble approaches

ABSTRACT

Over the last years, Internet spam content has spread enormously inside web sites mainly due to the emergence of new web technologies oriented towards the online sharing of resources and information. In such a situation, both academia and industry have shown their concern to accurately detect and effectively control web spam, resulting in a good number of anti-spam techniques currently available. However, the successful integration of different algorithms for web spam classification is still a challenge. In this context, the present study introduces WSF2, a novel web spam filtering framework specifically designed to take advantage of multiple classification schemes and algorithms. In detail, our approach encodes the life cycle of a case-based reasoning system, being able to use appropriate knowledge and dynamically adjust different parameters to ensure continuous improvement in filtering precision with the passage of time. In order to correctly evaluate the effectiveness of the dynamic model, we designed a set of experiments involving a publicly available corpus, as well as different simple well-known classifiers and ensemble approaches. The results revealed that WSF2 performed well, being able to take advantage of each classifier and to achieve a better performance when compared to other alternatives. WSF2 is an open-source project licensed under the terms of the LGPL publicly available at https://sourceforge.net/projects/wsf2c/.

1. Introduction and motivation

Today, the WWW (*World Wide Web*) is a key instrument to promote enterprises and organizations and advertise their products and services. However, the visibility of web sites is clearly influenced by the operation of the most used web search engines (e.g., Google, Yahoo Search, etc.). Therefore, during the last years SEO (*Search Engine Optimization*) techniques have gained popularity. In fact, more and more SEO advice and professional services are being demanded to meet promotional needs. At the same time that SEO has becomes indispensable to support legal e-commerce web sites, users selling illegal services and products have rediscovered SEO as a profitable tool set to disclose their services through the Internet, which has in turn originated the concept of Black Hat SEO, search spam or spamdexing (more commonly known as web spam).

As clearly stated in the work of Ghiam and Nemaney (2012), the capacity to detect and control web spam content is of foremost importance, essentially for the following reasons: (i) many spam pages are harmful for both search engines and the victim's (end user) machines, (ii) spam pages waste visitors' time and this may cause adverse effects on search engine results, and (iii) spam pages misuse important resources of search engines. This situation is aggravated by an increase in recent years of spam content inside web pages, mainly encouraged by the appearance of new web technologies oriented towards the online sharing of resources and information (e.g., cloud computing and/or social networks). Since it is a very complex and time consuming process for human experts to check all sites manually and determine whether they are spam or not, it has become essential to improve currently available web spam filtering platforms and techniques.

During the last twenty years, both Academia and IT Industry have fought against all forms of spam including junk e-mail (Biggio, Fumera, Pillai, & Roli, 2011; Guzzella & Caminhas, 2009; Pérez-Díaz, Ruano-Ordás, Fdez-Riverola, & Méndez, 2012), blog spam (Chu, Gianvecchio, Koehl, Wang, & Jajodia, 2013; Zhu, Sun, & Choi, 2011) or unsolicited SMS (Short Message Service)

* Corresponding author at: ESEI: Escuela Superior de Ingeniería Informática. Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004 Ourense, Spain. Tel.: +34 988 387013; fax: +34 988 387001.
E-mail addresses: jfgonzalez3@gmail.com (J. Fdez-Glez), drordas@uvigo.es (D. Ruano-Ordas), moncho.mendez@uvigo.es (J.R. Méndez), riverola@uvigo.es (F. Fdez-Riverola), rlaza@uvigo.es (R. Laza), pavon@uvigo.es (R. Pavón).

http://dx.doi.org/10.1016/j.eswa.2015.06.043
0957-4174/

(Delany, Buckley, & Greene, 2012). In the scientific context, there are many valuable contributions in the areas of feature selection (Forman, 2003; Shima, Todoriki, & Suzuki, 2004; Zorkadis & Karras, 2005) and classification techniques (Carpinter & Hunt, 2006; Cormack, 2006; Guzzella & Caminhas, 2009; Yu & Xu, 2008). Complementarily, from a commercial perspective, several e-mail anti-spam products based on the SpamAssassin[1] filtering operation have achieved great popularity due to their accuracy and the flexibility of SpamAssassin filters to integrate multiple classification techniques (Pecoraro, 2004; Symantec Corp., 2014). Moreover, in order to address the growth of spam deliveries, some authors have contributed important advances in SpamAssassin operation to address filtering throughput (Pérez-Díaz, Ruano-Ordás, Fdez-Riverola, and Méndez, 2013; Ruano-Ordás, Fdez-Glez, Fdez-Riverola, & Méndez, 2013).

However, and specifically related with the domain of web spam classification, we found a lack of probed methods and platforms able to successfully integrate and combine different techniques to increase the web spam filtering effectiveness. Only the work of Fdez-Glez et al. (2014) has shown how SpamAssassin operation can be successfully adapted to address the development of an integrative framework to filter web spam. Although this approach is theoretically able to effectively combine different filtering techniques to build up an accurate spam classifier, the underlying ensemble filter is unable to adapt itself to the dynamic nature of web spam contents, mainly because its rules and most of the combined classifier models are static.

In this context, and using CBR (*Case-Based Reasoning*) as an adequate methodology to design hybrid intelligent systems able to solve real life problems, the present study re-analyzes rule-based filtering systems and proposes a novel scheme to dynamically adjust filter threshold to continuously improve performance. In keeping with this goal, our model is able to effectively combine complementary weighted techniques (i.e., *experts*) to accurately classify different web sites by incorporating continuous updating of classifier capabilities. The operation mode of our approach is globally implemented by the sequential execution of four main stages: (i) retrieve the most appropriate filtering rules, (ii) reuse of these rules for assessing an initial solution, (iii) revise the confidence of the generated output and (iv) retain the system configuration parameters to maintain the filtering performance over time.

The structure of the paper is as follows: in Section 2 we summarize previous work carried out in the specific area of web spam filtering. In section 3 the proposed approach is presented in detail, covering the methods and knowledge used in its definition. Section 4 introduces the experimental framework and discusses the results obtained from different but complementary points of view. Finally, in Section 5 some concluding remarks are given and future work is outlined.

2. Related work in web spam filtering

The objective of this section is to provide a specific state of the art in web spam tricks as well as to characterize works and algorithms for web spam detection. The following two subsections discuss each of these aspects in detail.

2.1. Techniques and heuristics applied in spam attacks

Spammers use different spam techniques to fool current search engines (Spirin & Han, 2011). These techniques can be divided into four main categories according to the web spam taxonomy of Gyöngyi and Garcia-Molina (2005) and Najork (2009): content, link, page-hiding and click stratagems.

First of all, *content based spam* refers to any web spam technique that tries to improve the likelihood of a given page being returned as a search result by improving its ranking with the addition of salient keywords (Najork, 2009). Content based spamming techniques can be grouped based on the data field in which the spamming occurs (i.e., body, title, meta tag, anchor text or URL) or taking into consideration the type of terms that are added to the text fields (e.g., repetition of one or a few specific terms, dumping of a large number of unrelated terms, weaving of spam terms into copied contents or phrase stitching, etc.) (Gyöngyi & Garcia-Molina, 2005).

Link based spam is based on adding inappropriate and misleading associations between web pages to obtain a higher rank (Danandeh & Naser, 2014; Gyöngyi & Garcia-Molina, 2005). These kind of techniques can be grouped based on whether they add numerous *outgoing links* to popular pages (spammers often replicate some or all of the pages of a directory, and thus create massive outgoing link structures quickly) or they gather many *incoming links* to a single target page or group of pages (incoming links refer to any web spam technique that tries to increase the link-based score of a target web page by creating many hyperlinks pointing to it) (Najork, 2009).

Additionally, *page hiding based spam* presents a different content to search engines with the goal of obtaining a higher rank (Danandeh & Naser, 2014). The most common spam hiding techniques are content hiding, cloaking and redirection spam (Gyöngyi & Garcia-Molina, 2005). The first tactic makes spam terms or links on a page invisible when the browser shows its content. With cloaking, spam web servers return one specific HTML document to a regular web browser, while they send a different document to a web crawler. The third method of hiding the spam content on a page is by automatically redirecting the web browser to another URL as soon as the page is loaded.

Finally, *click spam* refers to the technique of submitting queries to search engines that retrieve target result pages and then "click" on these pages in order to simulate user interest in their content (Najork, 2009). Nowadays, result pages returned by leading search engines contain client-side scripts that report clicks on result URLs to the engine, which can then use this implicit relevance feedback for subsequent rankings (Spirin & Han, 2011).

2.2. Methods and algorithms for detecting web spam

Despite existing similarities with spam e-mail, specific research in this domain has attracted a fair number of scientists leading, in turn, to the development of novel approaches able to deal with the specific nature of web spam.

In this context, several techniques have focused on detecting *content-based web spam* by analyzing content features in pages (e.g., popular terms, topics, keywords or anchor text) to identify illegitimate changes. Among the early content spam papers, Fetterly, Manasse, and Najork (2004, 2005) statistically analyzed content properties of spam pages whereas Ntoulas, Najork, Manasse, and Fetterly (2006) used machine learning methods for detecting spam content. Later, a study by Erdélyi, Garzó, and Benczúr (2011) presented a comprehensive analysis of how various content features and machine-learning models can contribute to the quality of a web spam detection algorithm. As a result, successful classifiers were built using boosting, bagging and oversampling, in addition to feature selection (Geng, Jin, Zhang, & Zhang, 2013a; Nikulin, 2010). More recently, Prieto, Álvarez, López-García, and Cacheda (2012) presented a system called SAAD, in which web content is used to detect web spam.

In the context of *link spam*, PageRank and HITS methods were introduced by Page, Brin, Motwani, and Winograd (1998) and Kleinberg (1999), and are considered the first best solutions to

[1] The Apache SpamAssassin Project. Available at: http://spamassassin.apache.org/.

Please cite this article in press as: Fdez-Glez, J., et al. A dynamic model for integrating simple web spam classification techniques. *Expert Systems with Applications* (2015), http://dx.doi.org/10.1016/j.eswa.2015.06.043