Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 70 (2015) 434 – 441

Procedia Computer Science

4th International Conference on Eco-friendly Computing and Communication Systems (ICECCS)

Comprehensive Literature Review on Machine Learning structures for Web Spam Classification

Kwang Leng Goh[a], Ashutosh Kumar Singh[b]

[a]Curtin University, Kent St, Bentley WA 6102, Australia
[b]National Institute of Technology, Thanesar, Kurukshetra, 136119, India

Abstract

Various Web spam features and machine learning structures were constantly proposed to classify Web spam in recent years. The aim of this paper was to provide a comprehensive machine learning algorithms comparison within the Web spam detection community. Several machine learning algorithms and ensemble meta-algorithms as classifiers, area under receiver operating characteristic as performance evaluation and two public available datasets (WEBSPAM-UK2006 and WEBSPAM-UK2007) were experimented in this study. The results have shown that random forest with variations of AdaBoost had achieved 0.937 in WEBSPAM-UK2006 and 0.852 in WEBSPAM-UK2007.

Peer-review under responsibility of the Organizing Committee of ICECCS 2015

Keywords: Machine Learning ; Web Spam Classification ; Web Spamming

1. Introduction

In 2006, it was estimated that approximately one seventh of English webpages were spam, which became obstacles in users information-acquisition process[46]. In 2007, the cost of Web spam was estimated at US$ 100 billion globally and United States alone suffered an estimated cost of US$ 35 billion[4]. The intention of Web spam was to mislead search engines by boosting one page to undeserved rank. Consequently, it leaded Web user to irrelevant information. This kind of exploitation degraded the Web search engines by providing inappropriate or bias query results. Henzinger et al.[30] had identified Web spam as one of the most important challenges in Web search engine industries. Many people became frustrated by constantly finding spam sites when they were looking for legitimate content. In addition, Web spam had an economic impact since a high ranking provided large free advertising and so an increase in the Web traffic volume[3]. Even worse, at least 1.3% of all search queries directed to the Google search engine contain results that link to malicious pages[21]. In addition, one consultancy estimated that Russian spammers earned roughly

* Corresponding author: Kwang Leng Goh
E-mail address: alex.goh@curtin.edu.au

Peer-review under responsibility of the Organizing Committee of ICECCS 2015
doi:10.1016/j.procs.2015.10.069

US$2M to US$3M per year and one IBM representative claimed that a single spamming botnet was earning close to $2M per day[31]. Search engine companies generally employed human experts who specialized in detecting Web spam, constantly scanning the Web looking for spamming activities. However, the spam detection process often time-consuming, expensive and difficult to automate.

The development of an automatic Web spam detection system was an interesting problem as it concerned massive amounts of data to be analysed, the involvement of multi-dimensional attribute space with potentially hundreds or thousands of dimensions, and the extremely dynamic nature for novel spamming techniques that emerged continuously[44]. Often, large amount of Web spam pages were generated using machines by stitching together grammatically from a large collection of sentences[23]. Thus, machine learning method provided an ideal solution due to its adaptive ability to learn the underlying patterns for classifying spam and non-spam[22]. Machine learning approach can be divided into two categories —features and structures. The former depicted as the input used for classification while the latter defined the machine learning algorithm that was used for learning.

In this paper, the machine learning algorithms for Web spam detection were focused. C4.5 decision tree[39] (DT) and support vector machine[19](SVM) were two commonly used machine learning approaches among the adversarial information retrieval community. However, there were some evidences showing that SVM actually outperforms DT. Despite of that, researchers had shown that the outcome of SVM is easily manipulated in adversarial classification tasks like spam filtering[10]. Furthermore, recent papers[9,48] indicated that by injecting contaminated training data, the accuracy of the SVM will be significantly degraded. Previous studies had shown that multilayer perceptrons (MLP) neural network as an alternative Web spam classification tool[28] over SVM. However, there were still other popular machine learning algorithms within Web spam literatures that were not compared. Closest to this paper was a Web spam study reported by Silva et al.[43] who reported precision, recall and F measure in their study. In this paper, the area under the receiver operating characteristic curve (AUC) is used to evaluate the performance in Web spam detection for the reason that it did not depend on any threshold[22] like precision, recall and F-measure, and it aimed at measuring the performance of the prediction of spamicity[18].

This paper aims to provide a comprehensive machine learning approaches comparison within the Web Spam detection community using a standardized performance evaluation metric area under the receiver operating characteristic curve. In addition, several ensemble meta-learning algorithms such as boosting, bagging, rotation forest and stacking were included in the comparison to improve the classifier. Two well-known public available Web spam datasets WEBSPAM-UK2006[14] and WEBSPAM-UK2007[49] are used in this paper. Both datasets were downloaded from the Laboratory of Web Algorithmics, Universit degli Studi di Milano, with the support of the DELIS EU - FET research project. The former dataset was also used in part of a Web Spam Challenge in 2007[15,16] while the later dataset was used in Web Spam Challenge 2008[18].

The remainder of this paper is organized as follows. Related works available in the literatures are reported in Section 2, followed by descriptions of machine learning algorithms and meta-algorithms that are presented for comparison in Section 3. Section 4 describes the datasets, performance evaluation and parameters settings of the classifiers. Section 5 presents the results and discussion and lastly the conclusion in Section 6.

2. Related Work

In recent year, researchers in the adversarial information retrieval community had moved towards machine learning approach to detect Web spam. Actually the Web spam problem can be viewed as a classification problem. Machine learning constructed Web spam classifiers have shown positive results due to their adaptive ability to learn the underlying patterns for classifying spam and non-spam. The WEBSPAM-UK datasets have made a leap in Web spam community for using various machine learning models. In fact, previously there are few Web spam challenge series Web spam challenge track I[15], II[16] and III[18] which aim is to bring both machine learning and information retrieval community to solve the Web spam labelling problem.

Becchetti et al.[5] study several link-based metrics which include rank propagation for links and probabilistic counting to improve the Web spam detection techniques. Moreover, the authors conducted another similar research[7] which include more link-based metrics such as degree correlation and number of neighbours, and as a result the metrics achieve 80.4% detection rate with 1.1% false positive using DT with Boosting on WEBSPAM-UK2002 dataset. Besides link-based features, some researchers[37] propose several content-based features for Web spam detection. The