Information Sciences 232 (2013) 167–187

Contents lists available at SciVerse ScienceDirect

Information Sciences

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

Automatic seed set expansion for trust propagation based anti-spam algorithms

Xianchao Zhang [a,*], Wenxin Liang [a], Shaoping Zhu [a], Bo Han [b]

[a] School of Software, Dalian University of Technology, Economy and Technology Development Area, 116620 Dalian, China
[b] Department of Computer Science and Software Engineering, University of Melbourne, ICT 6.05, 111 Barry St. Carlton, VIC 3010, Australia

ARTICLE INFO

Article history:
Received 17 April 2011
Received in revised form 28 November 2012
Accepted 24 December 2012
Available online 8 January 2013

Keywords:
Search engine
Link analysis
Web spam
Seed expansion
Trust propagating

ABSTRACT

Seed sets are of significant importance to trust propagation based anti-spam algorithms, e.g., TrustRank. Conventional approaches require manual evaluation to construct a seed set, which restricts the seed set to be small in size, since it would cost too much and may even be impossible to construct a very large seed set manually. *The detrimental effect will be caused to the final ranking results by the small-sized seed sets.* Thus, it is desirable to automatically expand an initial seed set to a larger one. In this paper, we propose an automatic seed set expansion algorithm (ASE) *which enriches* a small seed set to a much larger one. The intuition behind ASE is that *if a page is recommended by a number of trustworthy pages, the page itself should be trustworthy as well.* Since links on the Web *can be considered as a tool for* conveying recommendation, we call links recommending the same page a joint recommendation link structure. *The joint recommendation link structures with large enough support degrees are employed by ASE algorithm to obtain new seeds.* It can be proved that using the joint recommendation link structure with a suitable support degree, the probability of selecting a spam page as a new seed *almost to* zero, thus the quality of the expanded seed set *can be* guaranteed. Experimental results on the WEBSPAM-UK2007 dataset show that with the same manual evaluation efforts, ASE can automatically obtain a lot of reputable seeds with very high quality, and significantly improves the performance of trust propagation based algorithms such as TrustRank and CPV (Computing Page Values).

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

People are using search engines to find useful information on the web, and usually they are only interested in *a few top-ranked result pages*. Since a higher ranking in searching results can bring more traffic to web sites and more profit to their owners, there is an economic incentive for manipulating search engine ranking results through unethical methods. This kind of manipulation, which attempts to trigger an unjustifiably favorable relevance or importance for some web pages, with respect to the page's true value, is called spamming [14], or spam. Spam degrades the search engines' retrieval quality, harms web sites/pages that should be highly ranked and weakens the users' trust for search engines [11,36]. *Therefore,* it has become so prevalent that every commercial search engine has to take measures to identify and remove spam [16].

Many anti-spam techniques have been proposed so far, and among them link-based semi-automatic algorithms that propagate experts' initial judgments over a set of seed pages to the entire web, e.g., TrustRank [15] *and CPV [35],* are the most

* Corresponding author. Tel.: +86 0411 87571515.
E-mail addresses: xczhang@dlut.edu.cn (X. Zhang), liang@computer.org (W. Liang), zhushaoping@gmail.com (S. Zhu), tq010or@gmail.com (B. Han).

0020-0255/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ins.2012.12.035

promising ones. These approaches are effective *because* both human efforts and computing power are exploited to deal with the complexity of the Web.

The seed set plays an important role in *this kind of algorithms. Manual evaluation process is usually used to construct a seed set [15],* which restricts the seed set to be small in size. The small-sized seed set can cause detrimental effect on the final ranking results. When the number of seeds is small, the top ranked results will be filled with seeds [17]; also, a small seed set is not sufficiently representative to cover different topics on the web*and hence causes*topic biases [31]. However, it seems impossible for search engine providers to get *such a large seed set only by means of manual evaluation for both economical and technical reasons.* Firstly, it costs too much to perform manual evaluation; secondly, it is difficult to find experts with good knowledge about the web, including the web graph and web snapshots [36]; thirdly, it is always time-consuming for experts to perform an evaluation, *and in the meanwhile* the seed sets need periodical refreshing since spamming tricks are adaptive and the web evolves rapidly; finally, it is possible for spammers to obtain a certain number of in-links from reputable sites, and these ruined sites should not be used as seeds, *and* cannot be discovered by humans before spam is detected.

The problems of small-sized seed sets have been *addressed* by several researchers [17,31]. It is revealed in [17] that a large seed set can achieve better performance than a small one. However, in [17] *only the sites in* .gov and .edu domains *were* selected as seeds to form a larger seed set, which brought serious domain bias problems to the ranking results. *Consequently,* it is desirable to expand a small manually selected seed set to a much larger one *by selecting potential seeds without domain bias.* However, to the best of our knowledge, no previous work has been done to *solve* this problem.

In this paper, we propose the ASE (Automatic Seed set Expansion) algorithm *for expanding* a small seed set to a much larger one. The intuition behind ASE is that if a page is recommended by a number of trustworthy pages, the page itself should be trustworthy as well. Since links on the Web can be considered as a tool for conveying recommendation, we call links recommending the same page a joint recommendation link structure. The joint recommendation link structures with large enough support degrees are employed by ASE algorithm to obtain new seeds. It can be proved that using the joint recommendation link structure with a suitable support degree, the probability of selecting a spam page as a new seed almost to zero, thus the quality of the expanded seed set can be guaranteed. The idea of using joint recommendation link structures to obtain new seeds was proposed in the preliminary version of this paper [33]. In this paper, the preliminary version is extended in the following aspects. Firstly, the accuracy of selecting new reputable seeds using the joint recommendation link structures is proved and the selection of threshold is discussed. Secondly, an effective initial seed set selection method for the ASE algorithm is proposed. Thirdly, the experimental section is enriched with more detailed settings and the CPV algorithm is added as a baseline method to show the effectiveness of the ASE method. Finally, more related works and a longer reference list are provided.

Overall, the contributions of this paper are as follows:

1. Through both analysis and experiments, possible drawbacks caused by small-sized seed sets are summarized and thus prove that a large set of less domain-biased seeds is of importance.
2. The Joint Recommendation Link Structure and Reputation Support Degree (RSD) metric are presented for efficiently to help expanding seeds, which are inspired by the fact that social interactions can be used to predict relationships between individuals [3,10].
3. The ASE algorithm is proposed to expand a small seed set to a much larger one with abundant reputable seeds and fewer false positives. ASE makes use of joint recommendation link structures to select pages with high enough RSD as potential seeds such that the probability that a selected page is spam approximates zero.
4. An initial seed set selection method incorporated with ASE is proposed. It simultaneously uses both PageRank and Inverse PageRank to ensure the quality and the coverage of the initial seeds, which is also a good choice for other seed-based algorithms.
5. The effectiveness of ASE is evaluated through experiments on WEBSPAM-UK2007 dataset [1], and the results show that ASE significantly improves the performances of typical seed-based anti-spam algorithms such as TrustRank [15] and CPV [35] in terms of both reputable site promotion and spam site demotion.

The rest of the paper is organized as follows. Related work is reviewed in Section 2. Preliminaries are introduced in Section 3. The importance of the size of seed set is addressed in Section 4. The ASE algorithm is presented in Section 5. Experimental results on the WEBSPAM-UK2007 dataset are shown and discussed in Section 6. Finally, Section 7 concludes this paper and outlines our future work.

2. Related work

Link-based spam attempts to create a link structure, i.e. a link farm, to take advantage of link-based ranking algorithms, such as PageRank [8] and HITS [19]. *How to detect Link-based spam is a serious problem to be solved since it can gain high rankings for target spam pages and is hard to detect.*

There are some graph based anti-spam algorithms. The work [37] builds a discrete analogue of classification regularization theory by defining discrete operators of gradient, divergence and Laplacian on directed graphs. The recent research [2] proposes another algorithm that follows regularization theory to learn to detect spam hosts or pages on the Web.

Download English Version:

https://daneshyari.com/en/article/393626

Download Persian Version:

https://daneshyari.com/article/393626

Daneshyari.com