International Journal of Natural and Engineering Sciences 7 (3): 52-55, 2013
ISSN: 1307-1149, E-ISSN: 2146-0086, www.nobel.gen.tr

Truncated TrustRank: Propagating Trust Based on Target Node

Mahsa POURBAFRANI[1*] Mehdi SHAJARI[1]

[1] Department of Computer Engineering and IT, Amirkabir University of Technology, Tehran, IRAN

[2] Department of Computer Engineering and IT, Amirkabir University of Technology, Tehran, IRAN

*Corresponding author:
E-mail: mahsa.bafrani@aut.ac.ir

Received: October 09, 2013
Accepted: December 11, 2013

Abstract

Regarding impact of the ranking of web pages on financial goals, web search became a very important issue in web. Subsequently, spam pages appear, which try to deceive search engines. In this study, a framework has been proposed for detecting spam pages. In this framework, at first statistical analysis on the large data set of web pages is done, after that, by using new hybrid algorithm, trust score for both spam and nonspam pages are calculated. Presented framework using some features to classify pages as spam or nonspam based on link structure of web. Based on classifier results in TrustRank algorithm, every node in the first level of trusted seed receive the deserve trust. Worthiness of a node is specified by classifier results.

Experiments have been done on sample of Farsi web graph with 800,000 nodes. By using proposed framework and propagating trust based on nature of target node, number of spam nodes which receive high trust scores is less than TrustRank algorithm.

Keywords: Web Spam, Web Graph, Trust Propagation, Dis-trust Propagation, Spam Detection

INTRODUCTION

Nowadays search engines become one of the most essential web applications which finding information without them is impossible for many. On the other hand, as these engines are so important, most of companies are trying to gain high rank in search engine results for financial purpose.

The term Web spam refers to pages and links on the Internet which are created with the purpose of misleading search engines. Search engine by adding keywords and indexing the sites, can return spam pages as search results. Another spamming technique is to construct a large number of bogus websites which refer to a single target. As the number of incoming links to a web page are used in search engine ranking algorithms, spammers are using this method to raise rank of their pages [1].

Although some of important algorithms like PageRank which are used in top search engines, ranking pages correct for English web pages, however, for Farsi web pages, search results include spam pages. Previous algorithms which using trust and dis-trust in order to gain the best result in detecting spam nodes, should be consider the structure of Farsi spam pages. Trust and dis-trust algorithms are contrast with tree important issues: first, only good and bad seed sets are used. Second, Trust and dis-rust are propagated without considering differences among target nodes. Third, there are many methods for splitting and accumulating trust in web graph to choose. Since spam nodes exist in real network, having a ranking method with the lowest error which presents every node it's deserve position in ranking process is essential [2]. Most of the spam detection methods have percentage of false positive. By using the spam detection techniques and combining them, search engines can improve their results, the impact of spam pages and false positive errors are reduced and the accuracy of spam detection is increased.

Our goal is to find the more efficient techniques than existing techniques. To achieve this goal we exploit of existing methods to improve quality of search results. This is followed by the specific features of previous algorithms to solve the problem of spam pages. In section two, a number of algorithms that have been used for this purpose are presented. The proposed method will be described in detail in Section three. In the fourth section, we will evaluate the proposed method and the conclusions and future works are described in the last section.

Related Work

Web can be modeled as a directed graph. PageRank algorithm is the best known algorithm for ranking web pages in the web graph. This algorithm simulates visits of users by random models. TrustRank algorithm is a modified version of PageRank algorithm, the algorithm starts to propagate the trust of seed set which judged by humans through all web pages [3]. After convergence, good pages receive high trust score and bad pages get lower trust score. TrustRank algorithm operates based on equation (1).

$$t = \alpha . M^T . t + (1 - \alpha) . s \qquad (1)$$

In equation (1), linear matrix M is the normalized version of the web graph, t is a trust vector in TrustRank algorithm, α is the decade factor, s is normalized version of trust vector for

seed set of good pages (S^+). For example, if $p \notin S^+$ then, $s(p) = 0$ and if $p \in S^+$ then $s(p) = 1/S^+$.

Anti-TrustRank is another ranking algorithm based on trust and dis-trust which is based on general principles of TrustRank algorithm [4]. In this algorithm dis-trust score from seed set of bad pages (spam pages) propagated through inverse links. After convergence, spam pages should earn higher dis-trust scores, while good pages should earn low dis-trust scores. Equation (2) is Anti-Trust Rank algorithm formula.

$$a = \alpha' . N^T . a + (1 - \alpha) . s' \qquad (2)$$

In equation (2), N is the normalized linear inverse of the web graph, a is dis-trust vector of Anti-TrustRank algorithm, α is the decade factor, S^- is normalized distrust vector of a bad seed set(S^-). For example, if $s'(p) = 0$ then,$p \notin S^-$ and if $s'(p) = 1/S^-$, then, $p \in S^-$.

In [5] Truncated PageRank is described as ranking function based on link structure. In this algorithm the importance of topological neighbors which is close to target node is reduced. [6] Shows that spam pages can be very sensitive to changes in PageRank decade factor. In this case, using Truncated PageRank, in addition to, change decade factor, all decline functionality in PageRank has been changed.

Directly, one way to reduce spam pages rank is using decline function which is reduce the impact of first level link of target page, such as figure 1.

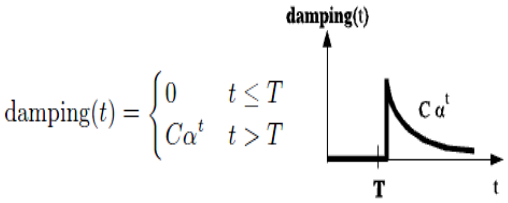

Figure 1. Damping factor of Truncated PageRank

In figure 1, C is a normalized constant and α is the decade factor used for PageRank. Reduction function penalized pages which gain most of their PageRank scores form first level links. This function is called reaction function of Truncated PageRank.

In [7], detecting spam pages is done only based on the links between pages. On the other hand, the web graph is intended regardless of content. Effective characteristics for ranking of web pages and probabilities of them are calculated in the web graph. These Features are used to build a classifier for detecting the network of spam links.

The Proposed Method

Considering advantages and disadvantages of spam detection methods which mentioned in previous section, we proposed a hybrid spam detection method. Proposed method is based on the combination of multiple algorithms to cover disadvantages of existing methods.

The overall structure of the proposed method is based on trust and dis-trust propagation in the web graph. In fact, the proposed method is based on link structure of web and link-based features. One of the important issues in trust and dis-trust propagation algorithms is competence of target pages.

In TrustRank algorithm, trust propagated from trusted seed without considering the nature of target pages. In other words, the nature of pages in the first level of the trusted pages (figure 2) should be declared. These first level nodes receiving highest trust score in web graph.

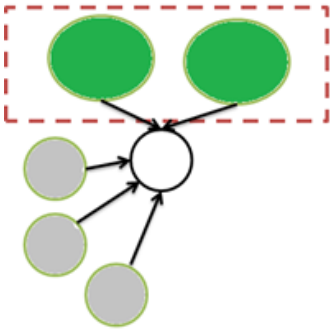

Figure 2. Impact of trust score of trusted seed on first level nodes

As aforementioned, in [8] most of reliable pages link to reliable pages. However, [1] has noted that some of the reliable pages link to spam pages. This problem is because some pages in a website being hacked by spammers or using spamming techniques such as putting spam pages link in the forums comments by them.

The TrustRank algorithm has two main stages, one for each parent, how to split trust between children called "Splitting" and another for each child, how to calculate earned points and trust scores from different parents called "Accumulation".

Practically, in the new method, one of changes to improve spam detection method is to change how each parent propagate and split their trust score to their children. In fact splitting step in previous methods has been replaced with new function to detect nature of children node.

We named our algorithm Truncated TrustRank which propagates trust based on nature of target node, this algorithm is shown in figure 3.

```
For n<N
  If l<L
    ClassifierOutput→Trust[n]
      If trust[n] =spam
        Dampingfacotor = alpha^l
              TrustRank[n] =
              dampingfactor*alpha*TrustRank[n] + (1-alpha)*TrustedSeed[n]
```

Figure 3. Truncated TrustRank algorithm

In above algorithm, N is number of node in web graph, L is the level of each node from trusted node (distance of target node from trusted parent), damping factor is factor to reduce trust score and alpha is similar to decade factor in original TrustRank algorithms. Therefore, in the proposed method, competence of target node in first level of trusted seed is considered. Competency of a node is determined using the implemented classifier. To implement the classifier, defined link-based features based on web pages structure and labeled pages with spam and nonspam label are used.

By running the algorithm on host graph instead of graph of pages another problem of trust and distrust propagation algorithms can be solved. The problem is each site can contains both spam and nonspam pages. Therefore, sites should be seen as a whole and their trust score should be the overall trust score of all pages. The evaluation result of the proposed solution is presented in next section.

EVALUATION And RESULTS

Experiments are done on real datasets collected from the Farsi web graph. By crawling web using the initial seed, Farsi web graph with 800,000 pages is created. After conversion of page graph to host graph, we have 25,000 sites or host in compressed format which is called BVGraph [9] that is ready for fast processing the graph.

Calculated Features for the English Web graph mentioned in [7]. Because of differences between structure of Farsi web graph and English web graph, effective features for detection spam sites in these two graphs are different.

The Truncated pagerank_1/PageRank, link exchange, in link degree, supporter1, supporter2 and node 'assortativity' are the most effective features for creating classifier in Farsi web graph.

By labeling 1000 node of a graph and using 50% of them to train and test, our classifier with precision 0.981 and false positive error rate 0.1 have been created. The classification algorithm has been used to implement the classifier is the J48 and the test and train method is based on 10-fold cross validation [10].

For choosing reduction factor measure in algorithm, some experiments have been carried out. This measure is largely based on the accuracy of implemented classifier. The 0.15 for this factor is the best value to reduce number of spam nodes in result.

Figure 4, shows the reduction of spam and nonspam host trust score in Truncated TrustRank than TrustRank algorithm. The horizontal axis shows three group which each one is the average value of reduction in score. In figure 4 (a) trust score of 70% of spam nodes reduce near to 0.06. While for nonspam figure 4 (b), more than 60% of them have 0.00001 changes in their trust scores. Figure 4, shows the improvement of TrustRank algorithm by reducing received trust of first level node of trusted seed based on their nature.

Figure 4, shows the reduction of spam and nonspam host trust score in Truncated TrustRank than TrustRank algorithm. The horizontal axis shows three group which each one is the average value of reduction in score. In figure 4 (a) trust score of 70% of spam nodes reduce near to 0.06. While for nonspam figure 4 (b), more than 60% of them have 0.00001 changes in their trust scores. Figure 4, shows the improvement of TrustRank algorithm by reducing received trust of first level node of trusted seed based on their nature.

For evaluating efficiency of Truncated TrustRank for spam detection a list of sites in decreasing order of their TrustRank scores are generated and divided to 10 segments which each buckets scores summing up to 10 percent of the total TrustRank score. The ratio of spam nodes to spam and nonspam nodes calculate using 50% of labeled hosts.

In figure 5, percentage of spam nodes to nonspam nodes in each bucket for TrustRank and Truncated TrustRank algorithm are shown. Based on segmentation, first bucket contain 1226 nodes which in result of Truncated TrustRank algorithm spam frequency is 1% while for TrustRank algorithm is near to 5%. The second buckets consists of 1856 nodes, which using our algorithm there is no spam while by using TrustRank algorithm 12% of nodes are spam.

In the same way, in the last buckets which are containing nodes with lower trust scores, percentage of spam nodes in Truncated TrustRank is more than TrustRank algorithm. In fact, spam nodes from first buckets have been transformed to the last buckets.

In some cases, similar to the nine categories, the proportion of spam to nonspam nodes in the two algorithms are close together, this can be because the proposed method is based on changing the nature of the nodes just in the first level of the trusted seed set. However, Truncated TrustRank algorithm shows improvement in performance of detecting spam.

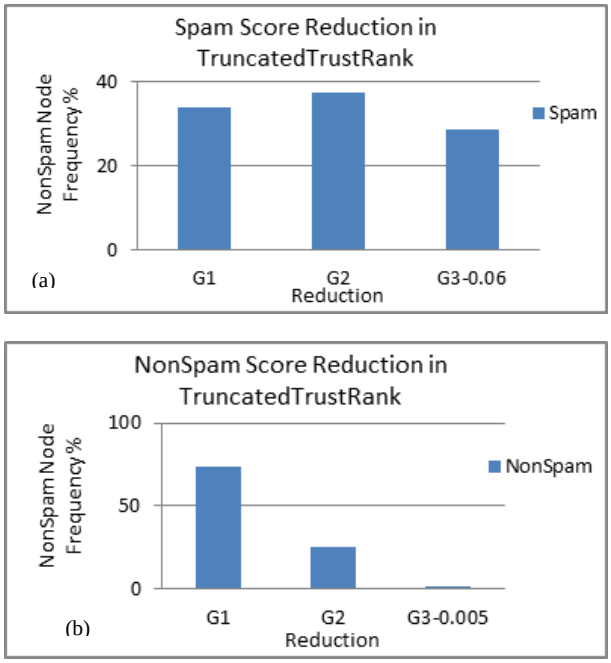

Figure 4. (a) Reduction of spam host trust score in Truncated TrustRank than TrustRank algorithm. (b) Reduction of nonspam host trust score in Truncated TrustRank than TrustRank algorithm

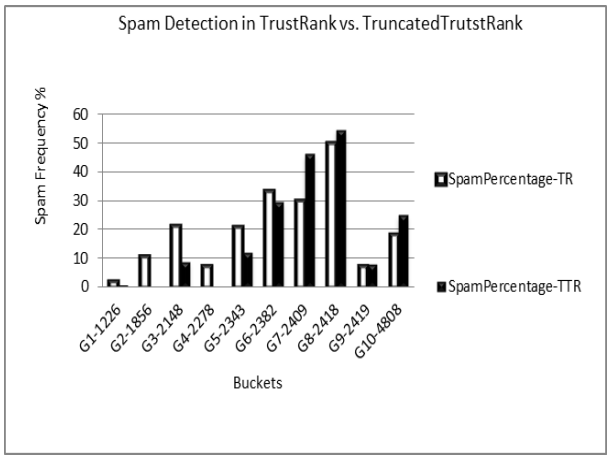

Figure 5. Spam Frequency of Truncated TrustRank versus TrustRank in 10 buckets

CONCLUSIONS

In this study, a framework has been proposed for detecting spam pages. In this framework, at first a statistical analysis on the large data set of web pages is done, after that by using a new hybrid algorithm trust score for both spam and nonspam pages are calculated. To this end, some of link-based metrics similar to the number of outgoing and incoming links is reviewed. Presented framework using these features for removing defects of TrustRank algorithm in lack of attention to nature of target node which receiving trust. By using this framework better result has been achieved in spam detection.

Experiments have been done on sample of Farsi web graph with 800000 nodes. Compared to TrustRank algorithm, results indicate by using proposed framework and propagating trust based on nature of target node, number of spam which is receive high trust score are reduced and more accuracy is achieved in detecting spam web pages.

REFERENCES

[1] Xianchao Zhang, You Wang, Nan Mou, Wenxin Liang, "Propagating Both Trust and Distrust withTarget Differentiation for Combating Web Spam", Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, 2011.

[2] Zolt´an Gyongyi, Hector Garcia-Molina,"Web Spam Taxonomy", First International Workshop on Adversarial Information Retrieval on the *Web*, May 2005.

[3] Vijay Krishnan, Rashmi Raj, "Web Spam Detection with Anti-Trust Rank", In the 2nd International Workshop on Adversarial Information Retrieval on the Web (AIRWeb), August 2006.

[4] Pruthi, Jyoti; Kumar, Ela, "Anti-Trust Rank: Fighting Web Spam", International Journal of Computer Science Issues (IJCSI), Jan2011, Vol. 8.

[5] L. Becchetti, C. Castillo, D. Donato, S. Leonardi, and R. Baeza-Yates. Using rank propagation and probabilistic counting for link-based spam detection. Technical report, DELIS-Dynamically Evolving, Large-Scale Information Systems, 2006.

[6] H. Zhang, A. Goel, R. Govindan, K. Mason, and B. Van Roy.Making eigenvector-based reputation systems robust to collusion. In Proceedings of the third Workshop on Web Graphs (WAW), volume 3243 of Lecture Notes in Computer Science, pages 92{104, Rome, Italy, October 2004. Springer.

[7] Luca Becchetti, Carlos Castillo, Debora Donato, Stefano Leonardi, Ricardo Baeza-Yates, "Using Rank Propagation and Probabilistic Counting for Link-Based Spam Detection ", 2006.

[8] Wu, B.; Goel, V, and Davison, B. D, "Topical trustrank: using topicality to combat web spam", In WWW '06, 2006.

[9] P. Boldi and S. Vigna. The webgraph framework I: compression techniques. In Proc. WWW, pages 595–602, 2004.

[10]Peter Mills (2011). "Efficient statistical classification of satellite measurements". International Journal of Remote Sensing.