International Journal of Information and Computation Technology.
ISSN 0974-2239 Volume 3, Number 4 (2013), pp. 241-250

http://www. irphouse.com /ijict.htm

Anti-Trust Rank for Detection of Web Spam and Seed Set Expansion

Ankit Srivastava[1] and Baisakhi Chakraborty[2]

Department of Information Technology, National Institute of Technology, Durgapur, West Bengal, INDIA.

Abstract

In the recent times, the Web has been the most popular and perhaps the most efficient platform for sharing, storing as well as retrieving information. Finding the required information from the Web is facilitated by search engines. Search engines form the interface between the Web and the users. Given the vast amount of information available on the Web, search engines must pick a small subset of the most relevant pages corresponding to the users' query and rank them in order to present the users a short list of high quality pages. Web spamming can significantly reduce the efficiency and reliability of search engines. Hence, there is an incentive for search engines to detect spamming efficiently and accurately. Human experts can easily detect spamming but to automate the same can be a very difficult task. But using manual effort to check web pages for spamming is practically not feasible. Here we study the method of propagation of anti-trust to combat web spam, which semi-automates the process of web spam detection through intervention of human expertise and an efficient algorithm to expand a small seed set selected by human experts automatically depending on the hyper-linking structure of the web, called Automatic Seed Set Expansion (ASE).

Keywords: Web Model; Page Rank; Anti Trust Rank; Automatic Seed Set Expansion.

1. Introduction

Search engines dominate the Web business. Some people try hard to trick search engines to fulfill this goal through techniques, like link farm [1], honey pot, etc. That means deliberate creation of more bad (SPAM) pages [2] to falsely boost the page rank of spam pages. Such pages, created to fool search engines are known as SPAMDEXING.

Search engines enable the users to access the information on the web. Since a higher ranking in searching results brings 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 pages true value is called Web Spamming. Hence, there is an incentive for search engines to detect spamming efficiently and accurately.

Many anti-spamming techniques [2,3,4] have been proposed so far, such as judgment propagation approaches that propagate experts' initial judgments of either trust or distrust over a set of seed pages or sites through hyperlinks to the entire Web. e.g. Trust Rank [2]. Trust Rank is an approach to find the trustworthiness of web pages. Initially, a seed set is chosen which is checked manually for spam and a handful of pages among the seed set is declared good and the rest as spam. Now trust or goodness is propagated from these pages declared as "good" by human experts to other pages depending on the hyper-linking structure of the Web. The Anti-Trust [3,4] approach is a slight variation of the Trust Rank approach which is used to detect spam pages so they can be filtered by search engines beforehand to produce high quality result.

2. Page Rank and Anti-trust Rank

2.1 Page Rank and Inverse Page Rank

PageRank [5] is a well known algorithm that uses link information to assign global importance scores to all pages on the web. The intuition behind PageRank is that a web page is important if several other important web pages point to it. Correspondingly, PageRank is based on a mutual reinforcement between pages: the importance of a page *influences* and *is influenced* by the importance of other pages. The PageRank score **r**(*p*) of a page *p* is defined as:

$$\mathbf{r}(p) = \alpha \cdot \sum_{q:(q,p)\in\mathcal{E}} \frac{\mathbf{r}(q)}{\omega(q)} + (1-\alpha)\cdot\frac{1}{N},$$

where α is the decay or damping factor.

The equivalent matrix form is:

$$\mathbf{r} = \alpha \cdot \mathbf{T} \cdot \mathbf{r} + (1-\alpha)\cdot\frac{1}{N}\cdot\mathbf{1}_N.$$

Hence, the score of some page p is a sum of two components: one part of the score comes from pages that point to p, and the other (static) part of the score is equal for all web pages.

It is important to note that while the regular PageRank algorithm assigns the same static score to each page, a biased PageRank version may break this rule. In the matrix equation,

$r = a \cdot T \cdot r + (1-a) \cdot d$, vector d is a static score distribution vector of arbitrary, non-negative entries summing up to one. Vector d can be used to assign a non-zero static score to a set of special pages only; the score of such special pages is then spread during the iterations to the pages they point to.

2.2 Anti-trust Rank

2.2.1 Intuition

This intuition behind the trust-rank is the same *approximate principle* [2], i.e it is rare for a good page to point to a bad page, that applies to trust-rank. This principle also implies that the pages pointing to spam pages are very likely to be spam pages themselves. The Trust Rank algorithm started with a seed set of trustworthy pages and propagated Trust along the outgoing links. Likewise, in Anti-Trust Rank algorithm, Anti-Trust is propagated in the reverse direction along incoming links, starting from a seed set of spam pages. A page is classified as a spam page if it has Anti-Trust Rank value more than a chosen threshold value. Alternatively, we could choose to merely return the top n pages based on Anti-Trust Rank which would be the n pages that are most likely to be spam, as per the algorithm.

2.2.2 Selecting the Seed Set of Spam Pages

We would want such a seed set of spam pages from which anti-trust could be propagated to as many pages as possible in a few numbers of hops. At the same time, we would also prefer if a seed set can enable us to detect spam pages having relatively high page ranks. In regard to this, choosing our seed set of spam pages from among those with high page rank satisfies both these objectives. Later we will see an effective technique of seed selection called *Automatic Seed Set Expansion [7].* Pages with higher page ranks are those from which a lot of high ranked pages can be reached in a small number of hops if we go backwards along the incoming links. Hence, this fulfills the first objective of fast and effective anti-trust propagation. Also, including high ranked pages in our seed set increases the chances of detecting spam pages with relatively higher ranks, as high ranked pages are pointed to by high ranked pages generally. So, we rank the pages according to the Pagerank method and include them in our seed set in decreasing order of their page ranks. This helps us to check the "goodness" of high ranked pages which appear at the top of the search engine result pages(SERP). Thus, we meet the above mentioned twin objectives.

2.2.3 The Anti-trust Algorithm

1. Obtain a seed set of spam pages checked by human experts. Arrange the spam pages in decreasing order of page ranks and include the first "n" pages in the seed set.

n=number of spam pages in the seed set

2. Compute I i.e. the Inverse Transition matrix.

Let I = web graph Inverse Transition matrix from the above mentioned web model.

3. Run the biased page rank [5] algorithm on the matrix I, with the seed set as the teleport set [6].
4. Rank the pages in descending order of page rank scores. This represents an ordering of pages based on estimated spam content. Alternatively, set a threshold value and declare all pages with scores greater than the threshold as spam.

3. Automatic Seed Expansion

Selecting the seed set lies at the core of the performance of trust propagation based anti-spamming algorithms, e.g., Trust Rank [2]. If the seed set is selected manually then the seed set is restricted to a small size as it is practically infeasible to construct a large size seed set manually. The small-sized seed set can adversely affect the final ranking of the search engine result pages (SERPs). So, we would like to expand the initially manually selected seed set to a large one by applying some automation mechanism. The Automatic seed set expansion (ASE) [7] is one such way to automatically expand the seed set through a joint recommendation link structure. Traditionally we used to employ human expertise to evaluate web pages for their goodness, which have limitations both in terms of quality and quantity. Both quality and quantity of seed sets are critical to the performance of these anti-spamming algorithms. The impact of quality is obvious since the misjudgment of an initial seed will be augmented through propagation and cause trouble. The quantity of the seed set is also an important issue. When the number of seeds is small, the top ranked results will be filled with seeds; also, a small seed set is not sufficiently representative to cover different topics on the web, and it will thus cause topic biases. Because of the above mentioned reasons, it is desirable to expand a small manually selected seed set to a much larger one by selecting potential qualified seeds. The automatic seed expansion algorithm (ASE) utilizes a joint recommendation link structure to select seeds for expansion.

3.1 The Impact of the Seed Set

Below is mentioned the possible drawbacks of a small-sized seed set selected manually by human experts.

Firstly, if the initial seed set size is small then the final result of the trust rank will be strongly biased towards the high ranked pages included in the initial seed set.

Despite propagating trust to other perhaps high quality pages, the effect is attenuated due to trust damping [2] as we move away from the nodes in the seed.

Secondly, if seeds contain pages mostly belonging to a certain domain or topic then the trust-rank result may be biased towards these topics, since pages within the same domain or of same topics are usually firmly connected. Pages from the large portion domain will certainly gain higher trust scores from the seed set than those from the small portion domain. For instance, if there are 10 reputable seeds from .com domain and 990 reputable seeds from .gov domain, pages from .gov domain will gain higher trust scores than those close to .com domain seeds, which is unfair for .com domain in ranking. Thus, the initial seeds that belong to different domains should be re-weighted according to their portions or be enriched by new seeds from other domains.

Finally, the performance of trust rank also deteriorates because of some possible links from reputable to spam pages. For example, in the commenting section of message boards or forums, spammers may leave links to spam pages to boost their ranks undeservingly. Thus the initial seed set may contain a spam page which is pointed to by a reputable page as the assumption that reputable pages point only to reputable pages is violated.

Given the above discussion, a large number of reputable, less domain-biased seeds are needed for algorithms like Trust Rank. But it costs too much to obtain them by manual evaluation. An automatic algorithm to expand the initial seed set with many reputable seeds to remedy the flaws mentioned above is desirable.

3.2 Intuition

The intuition behind ASE is that a page trusted by many trustworthy pages can be trusted. Although a spam page may also get in-links from reputable pages by some tricks, it is highly improbable for it to get many such links.

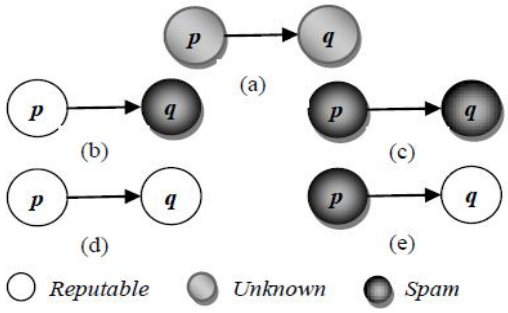

Figure 1: Directional Linkage Information.

Figure 1 describes several possible kinds of information we can get from two linked pages, where we use gray color to denote unknown pages, and white and black for reputable pages and spam pages, respectively. For a page *p* links to page *q*. (a) is

the common case. Though, we are unaware of either *p*'s or *q*'s quality, they must belong to one of four possible cases from (b) to (e).

Suppose *p* is unknown and *q* is reputable. Linking to a reputable page doesn't necessarily mean that page *p* is a reputable one. Thus, the target pair could in reality be either (d) or (e). However, if a page *q* is pointed to by reputable page *p*, it is more likely that *q* is reputable like (d), because links from reputable pages are more trustworthy than those from spam pages. There are certainly possible reputable-to-spam situations like (b), however, the probability of (b) is much less than for (d). This is because links on reputable pages intend to offer useful information by nature and rarely point to spam content. Besides, most reputable pages, especially those without message boards, etc., are not easy to be manipulated by spammers.

From the above discussion we can see that a page linked by a larger number of trustworthy pages can be trustworthy. Thus, if a certain number of reputable pages are known, more reputable ones can be identified by utilizing joint recommendation link structure information. In this way, a seed set of reputable pages can be expanded to a larger one.

3.3 Formulation of ASE

The notations used in the algorithm are as follows:

$V=V_L \ U \ V_{U,}$ where V denotes the node set of the web graph, and V_L and V_U represent the annotated page set and the unknown set , respectively. Specifically, for all pages in V_L, $V_L= V_R \, U \, V_S$, where V_R denotes the seed set of reputable pages and V_S denotes the spam page set in V_L, respectively.

The goal of Automatic Seed Expansion (ASE) is to select as large a number as possible of reputable pages V_E from V_U and add them to V_L. Before we discuss the ASE algorithm, Reputable Support Degree (*RSD*) is defined as a metric to measure the page's likelihood of being a reputable one.

Reputable Support Degree(RSD): *The reputable support degree of a page in the unknown set (denoted as RSD(p)) is the number of its in-links from annotated reputable pages.*

When the RSD of a page exceeds a certain predefined threshold value, that page is labeled as a new reputable page. Due to the motivation of spam as for gaining financial profits from the web, it is believed that commercial domain contains more spam than other domains, thus different reputable expansion threshold can be applied to create barriers for spammers in boosting commercial web pages.

In Figure 2, if thresholds for reputable government domain and commercial domain expansion are correspondingly 2 and 4, page *a* and *c* will be labeled as reputable pages, but *b* will be still unknown, since $N_a > T_g$, $N_b < T_c$ and $N_c > T_c$.

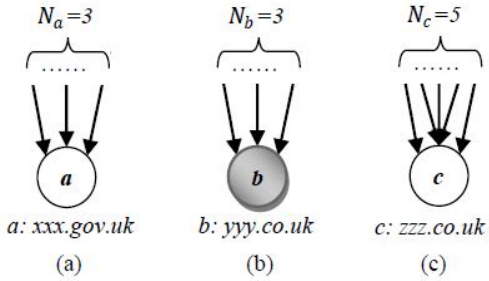

Figure 2: Domain specified threshold.

Sometimes, a spam page can also gather a certain number of reputable-to-spam links. Such kind of spam can be achieved for spammers by manipulated local graph topologies, e.g., to post contents with boosted page URL in blogs, forums, etc. To remedy the reputable-to-spam flaw, not all reputable pages are qualified to confer the authority during the expansion process. If a reputable page points to some spam pages in the seed set, we penalize the recommendation right of this page and drop its out-links, as it is hard for link-based algorithm to differentiate the manipulated links from qualified ones.

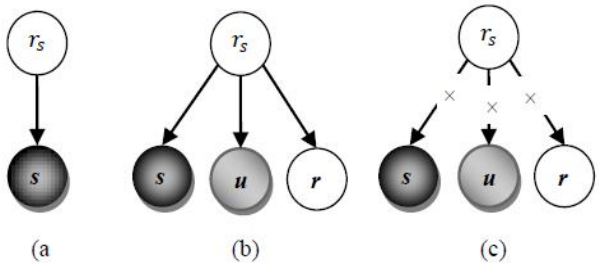

Figure 3: Reputable –to-spam link scenario.

In Figure 3, (a) is the typical scenario of the non-qualified reputable page rs, because it points to a spam page s. (b) reveals the general topology, the reputable page rs may points to many kinds of pages, including unknown, spam, and reputable pages. (c) is our method to alleviate such possible reputable-to-spam effects. We cut off the valuable support for spam page s, for potential reputable page r, though it loses support from reputable page rs, it can still obtain support from other robust reputable pages.

The above features are merged into the ASE algorithm and presented below, where RepList and ExpList are used to store the initial and temporarily added reputable seeds, respectively. And the spamNum is the counter of how many reputable links were received by the potential spam page. The reputable support vector vecRS for all unannotated pages is modified iteratively (step 5-12). If some items reach the expansion threshold, they will be labeled as new reputable ones and put them into ExpList (step 8-12). If a reputable page points to more than threshold number of spam pages, previous modification will be rolled back, and ASE will drop that seed's

authority (step 13-19). In the algorithm, each new reputable seeds' out-links will be scanned to update the reputable degree vector. If the reputable degree of an unknown node is over the corresponding threshold, move the node into the new expansion list. ASE run this operation iteratively until no more seeds can be added. Suppose there are totally n new expanded seeds and each one on average has k out-links. The time complexity is O(kn). In reality, because we can prune many seeds that have less than threshold in-links, n can be significantly reduced, therefore, the computation complexity is acceptable for real applications.

3.4 Automatic Seed Set Expansion Algorithm

Input: Adjacency matrix M, Threshold dictionary for different domains D, Initial manual evaluated seed set V_L, Unannotated seed set V_U.

Output: Expanded reputable seed set V_E.

1: Initialize $RepList = V_R$ (reputable seeds in V_L), reputable support vector $vecRS = \boldsymbol{0}$ for all unknown pages in V_U.

2: **while** *RepList* is NOT NULL

3: Initialize new temporary *ExpList*, *spamNum*.

4: **for** page p in *RepList*:

5: **for** page q in *OUT(p)*:

6: **if** q is in V_S (spam seeds in V_L):

7: $spamNum \leftarrow spamNum + 1$

8: **else if** q is in V_U:

9: $vecRS(q) \leftarrow vecRS(q) + 1$

10: **if** $vecRS(q) \geq D(domain(q))$:

11: Move q from V_U to V_E

12: *ExpList.add(q)*

13: **if** $spamNum \geq D(spam)$:

14: **for** page q in *OUT(p)*:

15: $vecRS(q) \leftarrow vecRS(q) - 1$

16: **if** q in *ExpList*:

17: Move q from V_E to V_U

18: *ExpList.remove(q)*

19: Drop *OUT(p)*

20: $RepList \leftarrow ExpList$

21: **return** V_E

4. Conclusion

The trustworthiness of web pages always seem to be in question, owing to evolution of spamming techniques with times and the spammers getting smarter and smarter. Trust propagation algorithms, such as TrustRank, have proved themselves quite effective in combating spam. However, the efficiency of these algorithms is restricted by the quality and the quantity of web pages in the seed set, since it is costly to obtain a large seed set via human intervention and evaluation. The ASE algorithm takes into account the hyperlink structure of the web to expand the initially manually determined seed set and grow it in size so that the biases towards high ranked pages as well as some domain based bias is significantly reduced . It improves the performance of Trust Rank significantly.

References

[1] Z. Gyongyi, Hector Garcia-Molina and Jan Pedersen(2004), Combating Web Spam with TrustRank, Proceedings of the 30th VLDB Conference, Toronto, Canada.

[2] Baoning Wu and Brian D. Davison (2005), Identifying Link Farm Spam Pages, Proceedings of the 14th International World Wide Web Conference, Chiba, Japan, pp. 820-829.

[3] Vijay Krishnan and Rashmi Raj (2006), Web Spam Detection with Anti-Trust Rank, AIRWeb, pp. 37-40.

[4] Jyoti Pruthi and Dr Ela Kumar (2011), Anti-Trust Rank: Fighting Web Spam, IJCSI International Journal of Computer Science Issues, Vol. 8,Issue 1.

[5] Sergey Brin and Lawrence Page (1998), The Anatomy of a Large-Scale Hypertextual Web Search Engine., Computer networks and ISDN Systems, pp. 107-117.

[6] Taher Haveliwala (2002), Topic-sensitive Page Rank, Proceedings of 11th. Conference on Worldwide Web, WWW2002, Honolulu, Hawaii, USA.

[7] Xianchao Zhang, Bo Han, Wenxin Liang (2013), Automatic Seed Set Expansion for Trust Propagation Based Anti-spamming Algorithms, Journal Information Sciences, Vol. 232, pp. 167-187.