Web Spam Detection with Anti-Trust Rank

Vijay Krishnan
Computer Science Department
Stanford University
Stanford, CA 4305
vijayk@cs.stanford.edu

Rashmi Raj
Computer Science Department
Stanford University
Stanford, CA 4305
rashmi@cs.stanford.edu

ABSTRACT

Spam pages on the web use various techniques to *artificially* achieve high rankings in search engine results. Human experts can do a good job of identifying spam pages and pages whose information is of dubious quality, but it is practically infeasible to use human effort for a large number of pages. Similar to the Trust Rank algorithm [1], we propose a method of selecting a seed set of pages to be evaluated by a human. We then use the link structure of the web and the manually labeled seed set, to detect other spam pages. Our experiments on the WebGraph dataset [3] show that our approach is very effective at detecting spam pages from a small seed set and achieves higher precision of spam page detection than the Trust Rank algorithm, apart from detecting pages with higher pageranks [10, 11], on an average.

1. INTRODUCTION

The term Web Spam refers to the pages that are created with the intention of misleading a search engine [1]. In order to put the tremendous amount of information on the web to use, search engines need to take into account the twin aspects of relevance and quality. The high commercial value associated with a web page appearing high on the search results of popular search engines, has led to several pages attempting spamdexing i.e. using various techniques to achieve higher-than-deserved rankings. Though it is not difficult for a human expert to recognize a spam web page, it is a challenging task to automate the same, since spammers are constantly coming up with more and more sophisticated techniques to beat search engines. Recent work [1], addressed the problem of Web Spam detection by exploiting the intuition that good pages i.e. those of high quality are very unlikely to point to spam pages or pages of low quality. They propagate *Trust* from the seed set of good pages recursively to the outgoing links. However, sometimes spam page creators manage to put a link to a spam page on a good page, for example by leaving their link on the comments section of a good page, or buy an expired domain. Thus, the trust propagation is *soft* and is designed to attenuate with distance. The Trust Rank approach thus starts with a seed set of trusted pages as the *teleport set* [2] and then runs a biased page-rank algorithm. The pages above a certain threshold are deemed trustworthy pages. If a page has a trust value below a chosen threshold value then it is marked as spam.

In our work, we exploit the same intuition, in a slightly different way. It follows from the intuition of [1] that it is also very unlikely for spam pages to be pointed to by good pages. Thus we start with a seed set of spam pages and propagate *Anti Trust* in the reverse direction with the objective of detecting the spam pages which can then be filtered by a search engine.

We find that on the task of finding spam pages with high precision, our approach outperforms Trust Rank. We also empirically found that the average page-rank of spam pages reported by Anti-Trust rank was typically much higher than those by Trust Rank. This is very advantageous because filtering of spam pages with high page-rank is a much bigger concern for search engines, as these pages are much more likely to be returned in response to user queries.

1.1 Our Contributions

- We introduce the Anti-Trust algorithm with an intuition similar to [1], for detecting untrustworthy pages.
- We show that it is possible to use a small seed set of manually labeled spam pages, and automatically detect several spam pages with high precision.
- We propose a method for selecting seed sets of pages to be manually labeled.
- We experimentally show that our method is very effective both at detecting spam pages as well as detecting spam pages with relatively high PageRanks.

2. ANTI-TRUST RANK

Our approach is broadly based on the same *approximate isolation* principle [1], i.e it is rare for a good page to point to a bad page. 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 our Anti-Trust Rank algorithm, *Anti-Trust* is propagated in the reverse direction along incoming links, starting from a seed set of spam pages. We could classify a page 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 our algorithm.

Interestingly, both Trust and Anti-Trust Rank approaches need not be used for something very specific like detecting

AIRWEB'06, August 10, 2006, Seattle, Washington, USA.

link spam alone. The *approximate isolation* principle can in general enable us to distinguish *good* pages from the not-so-good pages such as pages containing pornography and those selling cheap medication. Thus, for the purpose of our work we consider pages in the latter category as spam as well.

2.1 Selecting the Seed Set of Spam pages

We have similar concerns to [1], with regard to choosing a seed set of spam pages. We would like a seed set of pages from which *Anti-Trust* can be propagated to many pages with a small number of hops. We would also prefer if a seed set can enable us to detect spam pages having relatively high PageRanks. In our approach, choosing our seed set of spam pages from among those with high PageRank satisfies both these objectives.

Pages with high PageRank are those from which several pages can be reached in a few hops if we go backward along the incoming links. Thus this helps in our first objective. Also, having high PageRank pages in our seed set makes it somewhat more probable that the spam pages we detect would also have high PageRanks, since high PageRanks pages often get pointed to by other pages with high PageRank. We therefore select our seed set of spam pages from among the pages with high PageRank. This helps us nail our twin goals of fast reachability and detection of spam pages with high PageRank.

2.2 The Anti-Trust Algorithm

- Obtain a seed set of spam pages labeled by hand. Assign pages with high PageRanks for labeling by a human in order to get a seed set containing high PageRank pages.
- Compute T to be the Transpose of the binary web-graph matrix.
- Run the biased PageRank algorithm on the matrix T, with the seed set as the teleport set.
- Rank the pages in descending order of PageRank 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. EXPERIMENTS

3.1 Dataset

We ran our experiments on the WebGraph dataset, [3]. We chose data corresponding to a 2002 crawl of the "uk" domain containing about 18.5 millions nodes and 300 million links.

3.2 Evaluation Metric

Clearly, the only perfect way of evaluating our results is to manually check if the pages with high Anti-Trust score are indeed spam pages and vice-versa. It was observed in [1] that this process is very time consuming and often hard to do in practice.

We however circumvented this problem by coming up with a heuristic which in practice selects spam pages with nearly

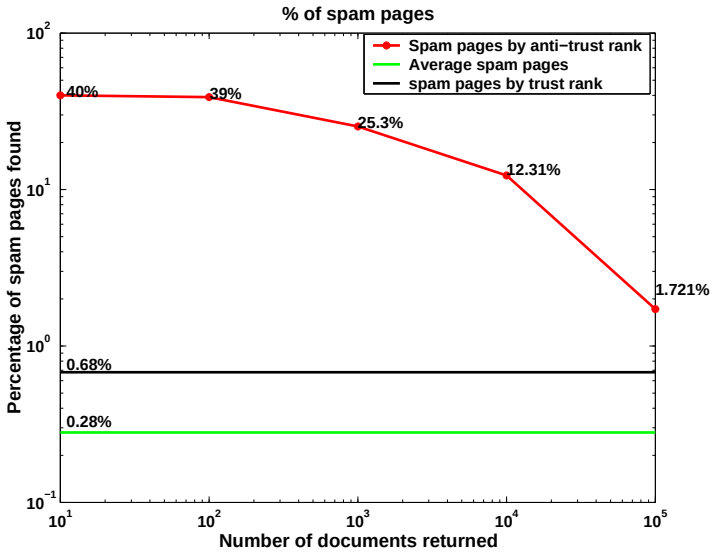

Figure 1: Comparison of the precisions of Anti-Trust Rank and Trust Rank at various levels of recall, against the naive baseline of total percentage of spam documents in the corpus. It can be seen what Anti-Trust Rank does significantly better than Trust Rank which is in turn clearly better than the naive baseline.

100% precision and also a recall which is a reasonable fraction of the set of true spam pages, on our dataset.

The Heuristic: We compiled a list of substrings whose presence in a URL almost certainly indicated that it was a spam page, on our dataset. As one would expect, our list contained strings like viagra, casino and hardporn. Thus, this heuristic enables us to measure the performance of our Anti-Trust Rank algorithm and compare it against the Trust Rank algorithm with a good degree of reliability. It seems reasonable to expect that the relative scores obtained by the spam detection algorithms with the evaluation being heuristic based would be representative of their actual performance in spam detection, since our heuristic has a pretty reasonable recall and is independent of both the Trust Rank and Anti-Trust Rank algorithms and would not give the algorithms we are looking at, an unfair advantage.

As per this heuristic, out of the 18,520,486 pages, 0.28 % i.e. 52,285 were spam pages.

3.3 Choosing the Seed Set

We chose the top 40 pages based on page rank from among the URLs that got flagged as spam by our heuristic. For comparing with Trust-Rank we picked the top 40 pages based on inverse page rank, among the pages marked non-spam by our heuristic. We also manually confirmed that the seed sets were indeed spam in the former cases and trustworthy pages in the latter case. We also studied the effect of increasing the seed set size in Anti-Trust rank. We found that we could benefit substantially from a larger seed set. Also we used the common α value of 0.85 i.e. the probability of teleporting to a seed node was 0.15.

3.4 Results and Analysis

From figure 1, we can see that both Anti-Trust Rank and

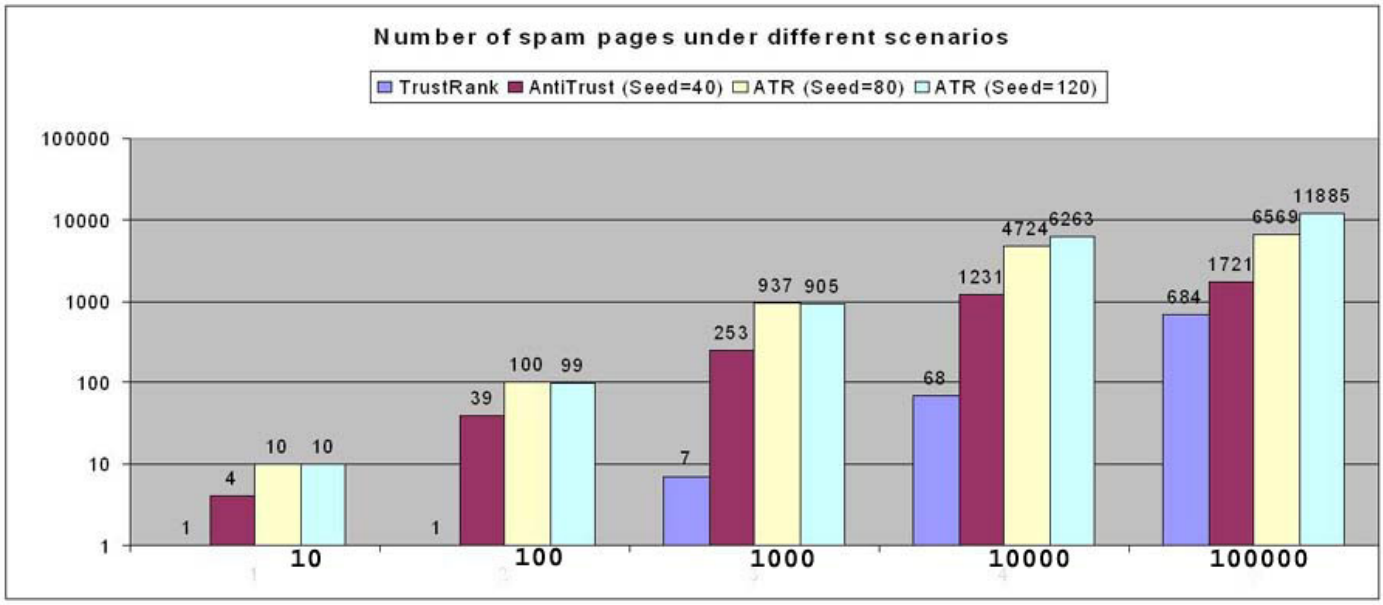

Figure 3: Comparison of the performance of Trust Rank with a seed set of 40 pages against Anti-Trust rank with 40, 80 and 120 pages respectively. The X-axis represents the number of documents selected having the highest Anti-Trust and lowest Trust scores. The Y-axis depicts, how many of those documents were actually spam(as measured by our heuristic). We observe that Anti-Trust rank typically has a much higher precision of reporting spam pages than Trust rank. Also, Anti-Trust rank benefits immensely with increasing seed-set size.

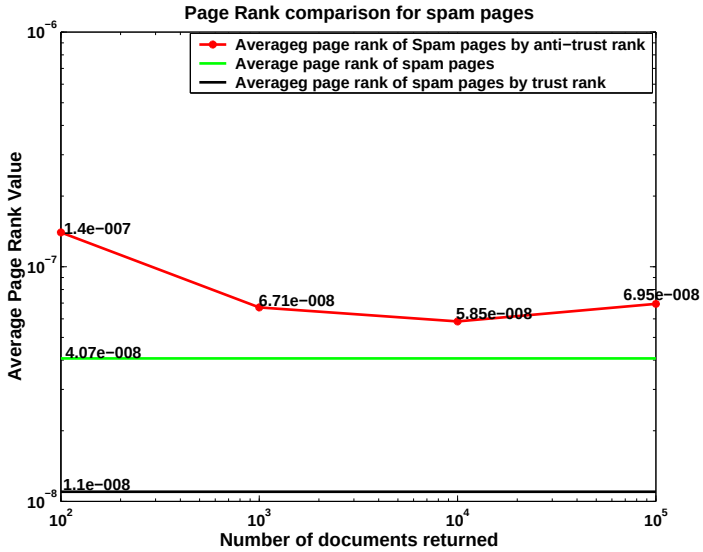

Figure 2: Comparison of the page ranks of spam pages returned by Anti-Trust Rank and Trust Rank at various levels of recall, against the baseline of average page rank of spam pages in the corpus. It can be seen that while Anti-Trust Rank returns spam pages with higher-than-average page ranks, Trust Rank returns spam pages with clearly lower-than-average page ranks.

Trust Rank are significantly better than the naive baseline corresponding to a random ordering of the pages, for which the precision of reporting spam would merely be the percentage of spam pages in the corpus. However we also see that Anti-Trust rank typically does much better than Trust Rank at different levels of recall.

This is intuitive because Trust Rank is capable of reporting with high confidence that the pages reachable in short paths from its seed set are trustworthy, while it cannot be expected to say anything with high confidence about pages that are far away from the seed set. Likewise our Anti-Trust Rank approach is capable of reporting with high confidence that the pages from which its seed set can be reached in short paths are untrustworthy.

Also, from figure 2, we find that the average rank of spam pages returned by Trust Rank is even lower than the average page rank of all spam pages. Anti-Trust rank however manages to return spam pages whose average page rank is substantially above the overall average page rank of all spam pages. The ratio of average page ranks of spam pages reported by Anti-Trust Rank and Trust rank was over 6:1 for different levels of recall. Thus, Anti-Trust rank has the added benefit of returning spam pages with high page rank, despite the fact that it has a significantly higher precision than Trust Rank at all levels of recall that we explored.

This is intuitive because, by starting with seed spam pages of high page rank, we would expect that walking backward would lead to a good number of spam pages of high page rank.

Figure 3 compares the performance of Trust Rank against Anti-Trust rank with an equal seed size of 40 and also show performance of Anti-Trust Rank with larger seed sets of 80 and 120 respectively. It shows the precisions achieved by

Trust Rank and Anti-Trust Rank at various levels of recall such as 10, 100, 1000, 10000 and 100000 web pages. We find that apart from achieving better precision of spam page detection than Trust Rank for the same seed set size, increasing the seed set size in Anti-Trust rank can lead to dramatic improvement in performance.

An analysis of success of these algorithms in picking trustworthy pages would not be very useful. This is because our corpus has over 99% trustworthy pages, and it would be very hard to conclude anything about the performance of these algorithms given that they would all attain a precision of well over 99% and would differ merely by a tiny fraction of a percent.

4. RELATED WORK

The BadRank algorithm [13] relies on intution similar to ours, namely that pages pointing to spam pages are likely to be spam themselves. The SpamRank algorithm [12] attempts to tackle link spam and assumes that spam pages have a biased distribution and attempts to compute the extent of underserved pagerank for a web page. The taxonomy of web spam has been well defined by [4]. There are many pieces of work on combating link spam. The problem of trust has also been studied in other distributed fields such as P2P systems [5]. Other approaches rely on detecting anomalies in statistics gathered through web crawls [7]. Approaches such as [8], focus on higher-level connectivity between sites and between top-level domains for identifying link spams. The data mining and web mining community has also worked on identifying link farms. Various farm structures and alliances that can impact ranking of a page has been studies by [6]. [9] identifies link farm spam pages by looking for certain patterns in the webgraph structure.

5. CONCLUSION AND FUTURE WORK

We have proposed the Anti-Trust Rank algorithm, and shown that it outperforms the Trust Rank algorithm at the task of detecting spam pages with high precision, at various levels of recall. Also, we show that our algorithm tends to detect spam pages with relatively high PageRanks, which is a very desirable objective.

It would be interesting to study the effect of combining these both the Trust Rank and Anti-Trust Rank methods especially on data containing a very high percentage of spam pages. It would also be interesting to attempt combining these link-based spam detection techniques with techniques that take text into account, such as text classifiers trained to detect spam pages.

Acknowledgements

We would like to thank Zoltán Gyöngyi, Dr. Anand Rajaraman and Dr. Jeffrey D. Ullman for helpful discussions. We would also like to express our gratitude to Paolo Boldi and Sebastiano Vigna whose compressed WebGraph dataset, with its useful Java API's made it very convenient for us to run experiments on a significant sized subgraph of the web.

6. REFERENCES

[1] Combating Web Spam with Trust Rank. Z. Gyöngyi, H. Garcia-Molina and J. Pedersen. Proc. of the 30st International Conference on Very Large Data Bases (VLDB), 2004.

[2] Topic-sensitive Page Rank. Taher Haveliwala. Proc. of the 11th International World Wide Web Conference, 2002.

[3] The WebGraph dataset. Online at: *http://webgraph-data.dsi.unimi.it/*

[4] Web Spam Taxonomy. Zoltán Gyöngyi, Hector Garcia-Molina. Proc. of the First International Workshop on Adversarial Information Retrieval on the Web (at the 14th International World Wide Web Conference), 2005.

[5] The EigenTrust algorithm for reputation management in P2P networks. S. Kamvar, M. Schlosser, and H. Garcia-Molina. Proc. of the Twelfth International World Wide Web Conference, 2003.

[6] Link Spam Alliances. Zoltán Gyöngyi, Hector Garcia-Molina. Proc. of the 31st International Conference on Very Large Data Bases (VLDB), 2005.

[7] Spam, Damn Spam, and Statistics. Dennis Fetterly, Mark Manasse and Marc Najork. Proc. of the Seventh International Workshop on the Web and Databases (WebDB 2004), 2004, Paris, France.

[8] Links to Whom: Mining Linkage between Web Sites. K. Bharat, B. Chang, M. Henzinger, and M. Ruhl. Proc. of the IEEE International Conference on Data Mining, 2001.

[9] Identifying Link Farm Spam Pages. Baoning Wu, Brian D. Davison. Proc. of the 14th International World Wide Web Conference, 2005.

[10] The PageRank citation ranking: Bringing order to the web. L. Page, S. Brin, R. Motwani and T. Winograd. Technical Report, Stanford University, 1998.

[11] The Anatomy of a Large-Scale Hypertextual Web Search Engine. Sergey Brin and Lawrence Page. Proc. of the 7th International World Wide Web Conference, 1998.

[12] SpamRank Fully Automatic Link Spam Detection. Andras A. Benczur, Karoly Csalogany, Tamas Sarlos, Mate Uher. Proc. of the First International Workshop on Adversarial Information Retrieval on the Web (at the 14th International World Wide Web Conference), Chiba, Japan, 2005.

[13] BadRank. Online at: *http://pr.efactory.de/e-pr0.shtml*