Extension and Propagation of Manual and Automatic Web Spam Scores

Tony Abou-Assaleh
GenieKnows.com
1567 Argyle Street
Halifax, Nova Scotia, B3J 2B2, Canada
taa@genieknows.com

Tapajyoti Das
GenieKnows.com
1567 Argyle Street
Halifax, Nova Scotia, B3J 2B2, Canada
tdas@genieknows.com

1. INTRODUCTION

We describe a Web spam detection algorithm that extends and propagates manual and automatic labels of Web hosts. The manual labels are derived from the training labels provided with the WEBSPAM-UK2006 dataset [1]. The automatic labelling assigned a spam label to hosts with a low variance in the out-degree of in-neighbours and hosts with significant overlap between their in-links and out-links. The score extension and propagation where applied to the directed host graph. Our approach achieves a 0.94 F1 score on the `webspam-uk2006-labels-DomainOrTwoHumans.txt` labels of 4948 normal and 674 spam labels, and a 0.93 F1 score on a our expanded labels of 5099 normal and 1618 spam labels.

Our extension and propagation algorithm is based on two assumption: 1) only spam pages link to spam pages, while normal pages do not; and 2) good pages link only to other good pages.

2. ALGORITHM

Our approach is divided into 7 steps described below.

1. *Identify the spam core set.* The spam core set consists of merging 3 sets: labelled spam, variance spam, and overlap spam. The labelled spam is the set of hosts from the `webspam-uk2006-labels.txt` file that include at least one human judgement as spam. The variance spam is the set of hosts that have a low variance in the out-degree of their in-neighbours. We used a variance threshold of 0.5. The overlap spam is the set of hosts that have an overlap of at least 5 between their in-links and out-links.

2. *Identify the normal core set.* the normal (nonspam) core set consists of the hosts labelled as normal by at least 2 human judges, and hosts from the `.ac.uk`, `.sch.uk`, `.gov.uk`, `.mod.uk`, `.nhs.uk` and `.police.uk` domains.

3. *Extend the spam core set.* We extend the spam core set by adding to the core set all the in-neighbours of domains in the labelled spam set. The motivation is that normal pages do not typically link to spam pages, but other spam pages do. We did not extend the variance spam and overlap spam sets because they include some false labels.

4. *Extend the normal core set.* We extend the normal core set by adding to the set all the out-neighbours of the core set. The motivation is that normal pages typically link only to normal pages.

5. *Initialize scores.* For each domain, we compute 3 scores: good, bad, and combined. We initialize the good score of extended normal set domains to +1, the bad score for extended spam set domains to -1, and all the other domain scores to 0.

6. *Propagate good and bad scores.* At each iteration, we propagate the good scores and the bad scores. The new good score for a host is calculated to be the discounted average of the good scores of the in-neighbours added to the hosts score. The discount factor is α^i, where i the iteration number and α is a number between 0 and 1. We use $\alpha = 0.2$ and a total of 10 iterations. After the last iteration, the combined score is computed as $\beta * bad + (1 - \beta) * good$, where β was set to 0.95.

7. *Assign labels to hosts.* Hosts with a negative combined score are assigned the spam label, while all other hosts are assigned the normal (nonspam) label. The scores where scaled and shifted to fall within the $[0, 1]$ range required by the submission format.

3. CONCLUSION

We present an algorithm for classifying Web sites as normal or spam. While many papers in the literature describe spam score propagation on undirected Web graph, we employ 2 assumption derived from empirical observation to propagate spam scores in a directed graph. Our method labels 3740 out of 11402 hosts as spam using all hosts, and 2511 out of 5780 hosts using the unlabelled hosts.

Future work includes a thorough analysis of the effect of the α, β, and the number of iteration on the performance of the algorithms. We are also investigating other high-precision statistical methods to be used in identifying the core spam and the core normal sets

Acknowledgement

We would like to extend our gratitude to the research team members at GenieKnows.com for their feedback and support.

4. REFERENCES

[1] C. Castillo, D. Donato, L. Becchetti, P. Boldi, M. Santini, and S. Vigna. A reference collection for web spam. *SIGIR Forum*, 40(2), December 2006.

AIRWeb '07, May 8, 2007 Banff, Alberta, Canada.