# Detecting Nepotistic Links by Language Model Disagreement[*]

András A. Benczúr  István Bíró  Károly Csalogány  Máté Uher

Computer and Automation Research Institute, Hungarian Academy of Sciences (MTA SZTAKI) and
Eötvös University, Budapest

{benczur, ibiro, cskaresz, umate}@ilab.sztaki.hu

## ABSTRACT

In this short note we demonstrate the applicability of hyperlink downweighting by means of language model disagreement. The method filters out hyperlinks with no relevance to the target page without the need of white and blacklists or human interaction. We fight various forms of nepotism such as common maintainers, ads, link exchanges or misused affiliate programs. Our method is tested on a 31 M page crawl of the `.de` domain with a manually classified 1000-page random sample.

**Categories and Subject Descriptors:** H.3.3 [Information Storage and Retrieval], I.7.5 [Document Capture]: Document analysis

**General Terms:** Algorithms, Measurement, Experimentation

**Keywords:** Language Modeling, Anchor Text, Link Spam

## 1. INTRODUCTION

Identifying and preventing spam is cited as one of the top challenges in web search engines. As all major search engines incorporate anchor text and link analysis algorithms into their ranking schemes, Web spam appears in sophisticated forms that manipulate content as well as linkage [5].

In this paper we concentrate on identifying hyperlinks between topically dissimilar pages. Our key result is the feasibility of the language model disagreement technique [7] for spam filtering in the scale of the entire Web, both in terms of algorithmic efficiency and quality. Mishne et al. [7] demonstrate that the distribution of words (a unigram language model) is a strong feature for telling legitimate and spam blog comments apart. We analyze inter-document relationship over the entire corpus by solving anchor text model comparison and prediction aggregation. We have similar goals as Davison [3] who trains a decision tree to distinguish navigational and link-spam links from the good ones. We target at penalizing links that are, in Davison's [3] terminology, nepotistic and "are present for reasons other than merit."

Links between topically unrelated pages may not necessarily be malicious; however they draw undeserved attention to the target. As examples, links to owners, maintainers, employee personal pages typically have no spamming intent but may have an effect of artificially ranking the target high. The widely investigated comment spam in blogs and guest books [7] form the malicious examples. Gyöngyi et al. [5] give more examples such as mirroring with the sole purpose of linkage to spam targets.

Our method fights a combination of link, content and anchor text spam. We catch link spam by penalizing certain hyperlinks and compute modified PageRank values as in [4, 6, 1, 2]. At the same time we also identify content spam if it has no trusted source of backlinks from the same topic. Finally we directly penalize false anchor hits that give very high value in Web IR systems, although measurements of this effect are beyond the scope of this report. We also remark the possibility to combine our method with link farm [4, 2] and navigational link [3] detection that detect different aspects of spamming and nepotism.

## 2. ALGORITHM

We present an algorithm that identifies hyperlinks where the language model of the target and the source disagree. We then feed suspicious edges into a weighted PageRank calculation [1] to obtain NRank, the "nepotism rank" of the page that we suggest be subtracted from the original PageRank values.

As in [7], our key ingredient is the Kullback-Leibler divergence (KL) between the unigram language model of the target and source pages. In fact it is infeasible to compute KL for all pairs of documents connected by hyperlinks. Two computationally easier tasks are to compare each *anchor text* to (i) the document containing it (as in [7]) and to (ii) the document pointed by it. While the former task is simply performed by a linear scan, the latter task requires an external memory sorting of all anchor text found.

We set aside the hyperlink if the corresponding language models differ. Since we assume that a typical anchor spam is generated by the owner of the page, we consider case (ii) above, complementary to the malicious anchors of reputable pages in [7]. We observe best performance when we extend the anchor text by a few neighboring words to properly handle very short anchor such as `"here"`; we obey segment boundaries defined by HTML and punctuation.

By using Interpolated Aggregate Smoothing as in [7], the language model for document $D$ has the form

$$p(w|D) = \lambda \frac{tf(w,D)}{\sum_{v\in D} tf(v,D)} + (1-\lambda)\frac{tf(w,C)}{\sum_{v\in C} tf(v,C)} \qquad (1)$$

where $C$ is the text of the entire corpus and $w$ is a word. We build a language model similar for an anchor $A$. In our experiments we set $\lambda = 0.8$; we smooth anchor term frequencies by the corpus formed by all extended anchor text. Finally we compute the Kullback-Leibler divergence

$$KL(A\,||D) = \sum_w p\,(w|A) \log \frac{p\,(w|A)}{p\,(w|D)}, \qquad (2)$$

a formula asymmetric in $A$ and $D$. The current form weights words by their relevance within anchors; we observed degradation in per-

9[*]Support from project NKFP-2/0024/2005, NKFP-2004 project Language Miner, ETIK and OTKA grants.

*WWW 2006,* May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

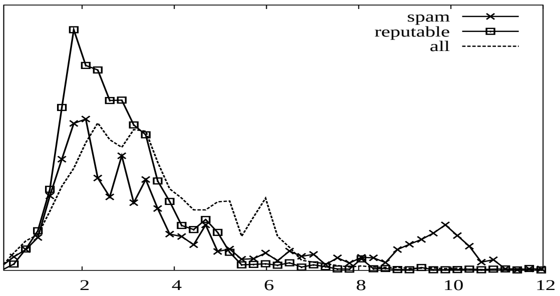

**Figure 1: Distribution of KL between anchor text and target document with our spam and reputable sample shown.**

formance when computing penalties by exchanging the role of $A$ and $D$ in (2).

As suggested in [7] the distribution of (2) is a mixture of Gaussians. KL will have normal distribution over the documents if all anchor text behave the same since we sum random variables that correspond to words and the words themselves have sufficient independence to yield a normally distributed sum. If however we have fair and malicious hyperlinks, the two categories will be biased towards the smaller and the larger values, respectively. While observations fit very well for case (i) anchor text and containing documents, for case (ii) anchor text and pointed documents behave more complex with spam taking lead around KL $\approx 4$ to 5 with a clear separating component with mean around 10, as seen in Fig. 1. The figure is based on the manually classified sample of [2].

In our algorithm we form the set of suspicious hyperlinks with KL value (2) above a threshold. We obtain NRank by feeding suspicious edges into PageRank by keeping edge above 7. Results are useful in the range 4-7; increased values of the threshold give NRank results farther from original PageRank and improve recall.

## 3. EXPERIMENTS

Torsten Suel and Yen-Yu Chen kindly provided us with a 31.2 M page April 2004 crawl of the `.de` domain. To evaluate the performance of our algorithms we used the manually classified stratified sample of [2]. The sample consists of 1000 pages and selected by first ordering the pages according to their PageRank value and assigning them to 20 consecutive buckets such that each bucket contained 5% of the total PageRank sum with bucket 1 containing the page with the highest PageRank. From each bucket 50 URLs are chosen uniformly at random, resulting in a 1000 page sample heavily biased toward pages with high PageRank which we manually classified into reputable and spam categories (see [6, 2] for details).

We measure the efficiency of our method by assiging each page to one of the 20 NRank buckets, the $i$th NRank bucket having exactly the same number of pages in it as the $i$th PageRank bucket. In Figure 2, top, we see that the top NRank buckets contain a very large amount of spam. And in Figure 2, bottom, we show how NRank distinguishes between spam and reputable pages by plotting the average difference between the PageRank and the NRank bucket number separately in each PageRank bucket. On the average we observe reputable pages have significantly larger demotion in NRank compared to PageRank than spam pages. We show pages of the `thema-⋆.de` click, a link farm with no useful content separate, as these pages use a simplistic but coherent e-commerce language. At low PageRanks legitimate pages are penalized slightly more than spam ones; notice however the real useful NRank penalties are never based on the bottom buckets. Also note that manual spam classification is particularly noisy at low qualities and the sample may also be less representative here.

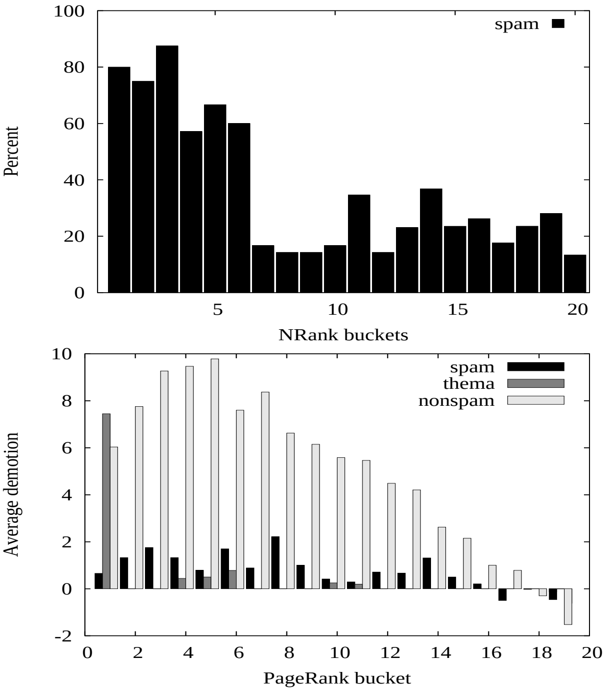

**Figure 2: Fraction of spam in NRank buckets (top) and average demotion of reputable and spam pages into NRank buckets as a function of their PageRank bucket (bottom).**

## 4. CONCLUSIONS

Our experiments show the applicability of language model disagreement along hyperlinks to differentiating among spam and nonspam pages. A number of questions left to subsequent work are as follows. Explore the effects of models and parameters (e.g. use $n$-gram models, smoothing, different penalty functions) and assess variants of the algorithm (e.g. by personalization). Measure the effect of NRank and anchor text downweighting on precision for popular or financially lucrative queries. Lastly evaluate the combination of content and link based spam filtering schemes.

## 5. ACKNOWLEDGEMENT

To Torsten Suel and Yen-Yu Chen for providing the `.de` web graph and Tamás Sarlós for many fruitful discussions.

## 6. REFERENCES

[1] R. Baeza-Yates and E. Davis. Web page ranking using link attributes. In *Proc. 13th International World Wide Web Conference (WWW)*, pages 328–329, 2004.

[2] A. A. Benczúr, K. Csalogány, T. Sarlós, and M. Uher. SpamRank – Fully automatic link spam detection. In *Proc. 1st AIRWeb*, 2005.

[3] B. D. Davison. Recognizing nepotistic links on the web. In *AAAI-2000 Workshop on Artificial Intelligence for Web Search*, pages 23–28, Austin, TX, 2000.

[4] N. Eiron, K. S. McCurley, and J. A. Tomlin. Ranking the web frontier. In *Proc. 13th International World Wide Web Conference (WWW)*, pages 309–318, 2004.

[5] Z. Gyöngyi and H. Garcia-Molina. Web spam taxonomy. In *Proc. 1st AIRWeb*, Chiba, Japan, 2005.

[6] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with TrustRank. In *Proc. 30th VLDB*, pages 576–587, Toronto, Canada, 2004.

[7] G. Mishne, D. Carmel, and R. Lempel. Blocking blog spam with language model disagreement. In *Proc. 1st AIRWeb*, Chiba, Japan, 2005.