Combining Anchor Text Categorization and Graph Analysis for Paid Link Detection

Kirill Nikolaev
Yandex
1 Samokatnaya str. Moscow.
7-495-739-70-00
kvn@yandex-team.ru

Ekaterina Zudina
Yandex
1 Samokatnaya str. Moscow.
7-495-739-70-00
zudina@yandex-team.ru

Andrey Gorshkov
Yandex
1 Samokatnaya str. Moscow.
7-495-739-70-00
gorshkov@yandex-team.ru

ABSTRACT

In order to artificially boost the rank of commercial pages in search engine results, search engine optimizers pay for links to these pages on other websites. Identifying paid links is important for a web search engine to produce highly relevant results. In this paper we introduce a novel method of identifying such links. We start with training a classifier of anchor text topics and analyzing web pages for diversity of their outgoing commercial links. Then we use this information and analyze link graph of the Russian Web to find pages that sell links and sites that buy links and to identify the paid links. Testing on manually marked samples showed high efficiency of the algorithm.

Categories and Subject Descriptors

H.3.3 [Information Search and Retrieval]: Information Filtering.

General Terms

Algorithms, Design, Experimentation.

Keywords

Search engines, language model, categorization, link analysis, machine learning, web mining.

1. INTRODUCTION

According to our recent observations the main method of the search engine optimization (SEO) in Russian Web is optimization via paid links. Though paid links do influence search engine ranking we do not treat them as spam links. Large part of paid links occurs in legitimate pages along with other useful links and often point to useful commercial sites. The paid links often cost significant money and that's why they are deliberately prepared. In this case, the anchor text always contains terms that match some commercial queries which are the target site keywords. Thousands of optimizers made this deliberate annotation by hand; therefore one may expect such links to contain useful information about their target. The ability to detect paid links helps to improve search engine ranking.

The work consists of two parts. The first part is text processing and topic classification, the second one is creating of a multi-topic page seed set and the link graph mark up using a modified HITS [1] algorithm hubs corresponding to link selling sites and authorities - to link-promoted sites. However, the main purpose of the algorithm is to detect every paid rather than link selling pages or paid link promoted sites.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

2. ALGORITHM

2.1 SEO-text Classifier

The parameter which indicates how "commercially interesting" a text fragment is will be called SEO-text score. From a popular SEO industry rating site we took a seed set of the SEO queries. On its basis we created an initial SEO text classifier similar to [2] in which only 2 topics, SEO and non-SEO, were used. Using the iteration method similar to the one we describe in 2.2 we got a large list of word unigrams (300 000) and bigrams (1 500 000) typical for SEO site link anchors. On the other hand a news text pool was used for construction of similar natural text unigrams and bigrams. Then we used this data to make an improved Bayes SEO-text classifier.

2.2 SEO-topic Classifier

To create an algorithm for SEO-topic identification we selected 22 topics most typical for commercial sites (for example: realty, finance, cargo carriage, etc). The algorithm to identify topics consists of 2 parts. We seeded a set of 3350 "mono-topic" words marked by hand; each word has its topic spectrum (TS). Then using many anchor texts with non-zero SEO-text score we calculated the TS for all other words according to the probability of a word's co-occurrence with words from the seed set in the same anchor text. This way we obtained 64 000 TS which were then used for anchor text categorization similar to [2].

At the second stage we used a simplified host-to-host link graph with 20 million edges containing non-zero SEO-text anchors. For every edge we identify two most probable topics using the aforesaid algorithm. We calculate the TS for the target vertexes using their incoming edges so that most of the targets have narrow spectra. For such targets we spread their topic upon all the incoming link anchors. On basis of these texts a new lexicon with about 200 000 words and 800 000 word pairs was created. The large number of terms lets us create a new efficient topic classifier based on the 1st order Markov chain [3].

The lexicon was manually adjusted according to the outrage mistakes analysis. Thus building such a huge lexicon needs rather little human effort. In fact we use automatically the work already done by optimizers.

2.3 SEO-out and SEO-in Classifiers

For further analysis we used a BHITS-like algorithm [4]. HITS and its various modifications were already used to find spam links [5] [6], now we use it to find paid links. We used a bipartite link graph (the source pages to the left and the target hosts to the right) with all known spam pages and links from link farms, etc. removed. We improved the standard HITS link preparation and deleted all links within one owner (owner is a second level domain which is not a hosting or a third level domain if it's located on a hosting). Thus we got a link graph with 300 million edges, 50 million source pages and 19 million target sites. Using the topic classifier (2.2) for the graph edges we got 1 million "mono-topic" targets.

In our algorithm we use the concepts of SEO-out and SEO-in scores that are analogues to the hub and authority scores in the classical HITS algorithm respectively. The SEO-out score shows the probability of the page being a link seller. The SEO-in score shows the probability of the site promoting itself with paid links. The sites with high SEO-in scores most are commercial resources which use expensive promotion to move up in the SERP.

A page pointing to different topic targets is a high probable link seller. A set of so defined multi-topic pages with high outgoing link SEO-text score and some other parameters was used as a page seed set (3 million pages). SEO-out and SEO-in scores are calculated according to standard HITS algorithm (fig. 1) using 2 iterations. At this stage our goal is to get a list of high SEO-in score target sites and finally we get about 500 000 of them.

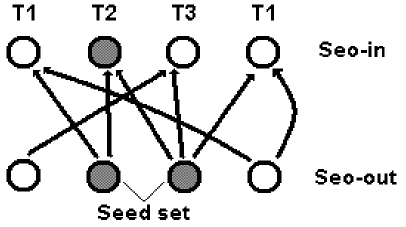

Figure 1. Calculating SEO-in scores according to SEO-out scores of multi-topic seed set pages in the bipartite link graph via HITS algorithm (1 iteration is shown, T1, T2, T3 are the target site topics).

2.4 SEO-link Classifier

We define the SEO-link score for a link as the probability of being a paid link. A simple one-pass algorithm computes this score for each link as follows. First we estimate the probability for the page to contain paid links (SEOout of the page) aggregating the following parameters: average target SEO-in score (AvgSEOin), average anchor SEO-text score (AvgSEOtext), number of the target topics (NTh) and also some other page clues according to the following formula:

$$SEOout = k_1 \times AvgSEOin + k_2 \times AvgSEOtext + k_3 \times NTh + \ldots \quad (1)$$

Then using these factors (anchor SEO-text, source page SEO-out, target site SEO-in and some other on-page clues about the link) we compute the desired SEO-link score as:

$$SEOlink = l_1 \times SEOtext + l_2 \times SEOin + l_3 \times SEOout + \ldots \quad (2)$$

The k_i and l_i factors were selected on a test learning sample of 2500 links marked by hand and about 10 000 links taken partly from Wikipedia and partly from known link selling pages.

This computation doesn't take much time and memory and can be executed with the link base processing program.

3. RESULTS

To estimate the precision and recall of the algorithms we used test samples marked up by 8 experts.

For the categorization algorithm evaluation we took top 2 200 sites related to our topics (top 100 for each topic) from a popular SEO industry rating site and selected randomly non-zero SEO-text incoming anchors. If the anchor topic was clear, the experts assigned the anchor one of 22 topics. One part of the sample (12 100 anchors) was used for testing and adjusting. The other part (3 800 anchors) was used for evaluation; 94% precision and 97% recall were estimated.

For paid link detection evaluation we used 2 samples (Tab.1). The 1st sample contains about 1700 useful natural and 1850 paid links taken randomly from the index and marked by hand (precision data was estimated on the natural link set only). We were able to identify links from one link exchange service directly. Thus we could get a collection of definitely paid links, which was used as 2nd sample.

From the whole amount of 300 million links in our graph 50 million were marked by our algorithm as paid links (17%).

Table 1. The paid link detection results.

Sample	Precision	Recall
1. 3 550 links	95%	93%
2. about 140 000 links	-	96%

4. CONCLUSION

Using this paid link marking up, the link relevance factors can be calculated differently for commercial and non-commercial queries. In the first case, paid links are taken into account and used specifically to improve the commercial ranking, while in the second case they aren't taken into account. This lets the ranking formula improve the search quality, decrease the over-optimization influence for non-commercial queries and increase the necessary SERP variety.

This algorithm can be improved by using the Yandex's page segmentor applying the microHITS analogue for link blocks [7].

5. ACKNOWLEDGMENTS

We would like to thank Sergey Pevtsov, Ilya Segalovich, Arkady Borkovsky and Sergey Volkov for fruitful discussions.

6. REFERENCES

[1] Kleinberg, J. (1997). Authoritative sources in a hyperlinked environment. Journal of the ACM 46 (5): 604–632.

[2] T. H. Haveliwala. Topic-sensitive pagerank. In Proc. 11th International WWW Conference, pages 517-526, 2002.

[3] Lafferty J., Zhai, C. Document language models, query models, and risk minimization for IR. In Proceedings of SIGIR-2001, pp 111-119.

[4] K. Bharat and M.R. Henzinger, Improved algorithms for topic distillation in a hyperlinked environment, Proc. 21st Annual International ACM SIGIR, pp.104–111, 1998.

[5] B. Wu and B. Davison. Undue influence: Eliminating the impact of link plagiarism on web search rankings. Technical report, LeHigh University, 2005.

[6] Yasuhito Asano, Yu Tezuka, Takao Nishizeki. Improvement of HITS algorithms for spam links. APWeb/WAIM 2007, LNCS 4505, pp 479-490, 2007.

[7] S. Chakrabarti. Integrating the Document Object Model with Hyperlinks for Enhanced Topic Distillation and Information Extraction. ACM 1-58113-348-0/01/0005, 2001.