Russian Web Spam Evolution: Yandex Experience

Sergey Pevtsov
Yandex, LLC
16 Leo Tolstoy st.
Moscow, Russia
spev@yandex-team.ru

Sergey Volkov
Yandex, LLC
16 Leo Tolstoy st.
Moscow, Russia
svolf@yandex-team.ru

ABSTRACT

Web spam has a negative impact on the search quality and users' satisfaction and forces search engines to waste resources to crawl, index, and rank it. Thus search engines are compelled to make significant efforts in order to fight web spam. Traffic from search engines plays a great role in online economics. It causes a tough competition for high positions in search results and increases the motivation of spammers to invent new spam techniques. At the same time, ranking algorithms become more complicated, as well as web spam detection methods. So, web spam constantly evolves which makes the problem of web spam detection always relevant and challenging. As the most popular search engine in Russia Yandex faces the problem of web spam and has some expertise in this matter. This article describes our experience in detection different types of web spam based on content, links, clicks, and user behavior. We also review aggressive advertising and fraud because they affect the user experience. Besides, we demonstrate the connection between classic web spam and modern social engineering approaches in fraud.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information search and retrieval

Keywords

Web Spam Detection, User Behaviour, Aggressive Advertizing

1. INTRODUCTION

Since search engines (SE) from the very start use web page content as a main source of ranking signal, content spam became one of the most widespread type of spam. Spammers try to impact the ranking factors like TF/IDF or BM25 by creating texts (often meaningless) with optimal keywords frequency. The detection of this type of spam is mostly tackled by language model based approaches [4], [8].

As soon as search engines started using link-based ranking features like Page Rank, spammers tried to impact these factors with link-farms, nepotistic links, paid links and other artificial linking formations. Most approaches to detect link spam are based on manifold web graph properties [1], [11].

Applying user behavior data for ranking causes emergence of corresponding spam techniques like click fraud. There are some algorithms to detect user behavior based spam [6], [9].

From our experience we know that any factor used in ranking can be exploited in order to get a better ranking position. Spammers are strongly motivated to find and exploit such features. Thereby search engines have to be always aware of these intentions when designing and implementing new features and algorithms.

Cheating with ranking features is not the only spammers' strategy to make money. Trying to increase traffic monetization many of them do not limit themselves by cheating search engines algorithms and start cheating users, so search engines have to take this into account. The social aspect of fighting web spam is not less important than the technical one.

This article is organized as follows. Section 1 describes Yandex's antispam infrastructure allowing to quickly respond to appearing threats. In the second section we briefly describe major types of web spam and demonstrate how search engines react to this phenomena. Section 3 is dedicated to aggressive advertising problem related to users' experience and satisfaction. The fourth section is about more and more evident close connection between web spam, malware, and social engineering fraud. Conclusions and unresolved burning problems are presented in Section 5.

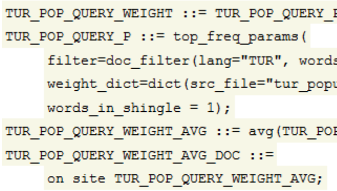
2. ANTISPAM INFRASTRUCTURE

There is a huge number of web pages in the Internet, so search engines need to process documents very fast. Yandex stores more than 20 bln. documents in its search index and crawls more than 3 bln. documents every day. Some documents are subject to special processing with JavaScript interpretation and full content rendering. We designed and implemented a special regular expressions-based language for building various content classifiers. This language is used in fast classification system that can process 200,000 documents per second. Several classifiers with high precision-recall characteristics were developed using these technologies. Content classifiers are based on rules which can be written easily by an analyst without strong programming skills (see Figure 1).

These rules serve as weak classifiers which are then fed as features to the gradient boosted decision trees [5] based web spam detection algorithms. There is huge amount of collected data related to domains, texts, links, clicks, anonymized user's behaviour, etc. that can be used to develop features. Apart from using the continuously updated set of labeled

Copyright is held by the International World Wide Web Conference Committee (IW3C2). IW3C2 reserves the right to provide a hyperlink to the author's site if the Material is used in electronic media.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.



```
TUR_POP_QUERY_WEIGHT ::= TUR_POP_QUERY_P.total_weight;
TUR_POP_QUERY_P ::= top_freq_params(
     filter=doc_filter(lang="TUR", words_in_doc>=100, tag="p"),
     weight_dict=dict(src_file="tur_popular_queries.dict"),
     words_in_shingle = 1);
TUR_POP_QUERY_WEIGHT_AVG ::= avg(TUR_POP_QUERY_WEIGHT);
TUR_POP_QUERY_WEIGHT_AVG_DOC ::=
     on site TUR_POP_QUERY_WEIGHT_AVG;
```

Figure 1: TUR_POP_QUERY_WEIGHT_AVG_DOC - document feature "average weight in the dictionary of popular turkish queries"

data prepared by a group of professional assessors, our algorithm is trained on the feedback that our technical support receives and processes every day. Such feedback allows to be timely aware of new trends in spam techniques and other types of frauds.

3. WEB SPAM FORMS

The main goal of webspam is to attract traffic from search engines. There are different strategies to maximize the number of visitors. Spammers take into consideration different factors like query popularity, competition level, ranking function used for certain types of queries, etc. It's worth mentioning that web spam is very close to search engine optimization (SEO). Certainly, there are lawful SEO practices ("white hat" SEO) where a web site is being analyzed to fix incorrect indexing and other technical problems. Unfortunately, other SEO techniques are used to aggressively promote sites in search engine results for a selected set of queries ("gray hat" and "black hat" SEO). There are 4 primary targets being used in site promotion: texts, domains, links, and user behaviour. Each target corresponds to a group of ranking features. "Black hat" and "gray hat" SEO are trying to get optimal feature values that result in the highest rank.

Further in the article we proceed from the most trivial spam techniques to more complicated. We describe the evolution of webspam and SEO from the technical point of view as well as discuss its social impact.

3.1 Texts

The trivial "optimization" is based on TF/IDF cheating and consists of experimenting with terms, their frequencies and placement on a page by trial-and-error methods. Such artificial term placement often makes page content worse and significantly spoil the user experience. Thus excessive text optimization should be controlled by SE and text ranking features require careful tuning.

3.2 Domain names

Effect of high weight of domain features is shown below. The simplest trick is using a domain with the name that contains words from a query being promoted. It might look odd, but usage of this technique to get profit easily makes all results for specific query look indistinguishable (each domain name in SERP's top results contains a substring from the query).

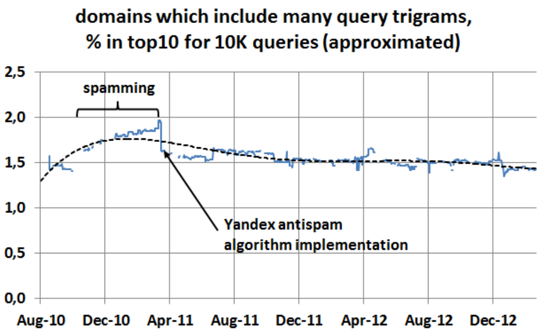


Figure 2: Spamming domain names in 2010 was intense but short

3.3 Links

Buying links at link-broker services is one the most popular link spamming technique in Russia. Unfortunately this practice is a quite common due to delay in search engines' reaction. Relatively long period of time when paid links were an effective way to promote sites allowed this technique to spread a lot. It demonstrates the urgent need for SE to react to new forms of spam very quickly. We believe that fast reaction of search engine to a new type of spam is often more important than the algorithm recall.

We developed effective algorithm that combines anchor text categorization and graph analysis to detect paid links [7]. Paid link classifier helps to calculate link relevance factors for commercial and non-commercial queries differently. This allows to improve ranking algorithm for better search quality, decrease SEO's influence for non-commercial queries and increase SERP diversity. It's important to notice that paid links are easy to classify which allows SE to control search quality and fight spam focused on commercial queries. But it's important to say that paid link phenomenon is firmly rooted in Russian segment of web due to slow reaction.

As a consequence, in the middle of 2007 all popular search engines in Russia were under strong SEO pressure[1]. It was relatively easy to promote websites into top-10 search results those days. Ranking algorithms were not proof to such cheating techniques and many common queries contained different intents were affected by link spam. For example 8 out of 10 results for "water" query included water delivery offers, 9 of 10 results for "insects" were about insects extermination, etc. As a result, many popular queries that do not have pronounced and unambiguous commercial intent were spammed by SEO and search results for them consisted of commercial offers mostly, that decreased search quality and diversity. Over the next years search engines improved their algorithms, made sites promotion much more complicated and significantly increased results diversity (Figure 3).

3.4 User behaviour

Since user behavior is also good source of ranking signal [2], spammers started to investigate how to exploit it. They do not know how click features are calculated and work, but reasonably assume that CTR does play important role. The most trivial technique used is to find a website in SE and to start clicking on it. The next step is straightforward: to create a community using Pay-Per-Action model where low cost workers perform tasks of querying SE and clicking

[1] according to automatic independent search quality analytic project analyzethis.ru

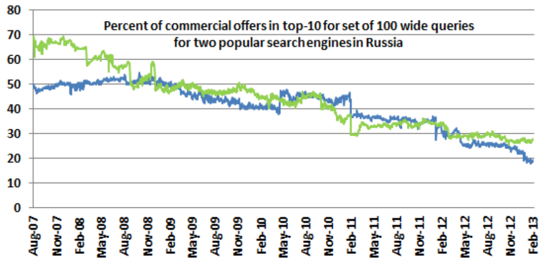


Figure 3: Percent of commercial offers in top decreased, search results diversity on transactional queries significantly increased (see `analyzethis.ru`)

results according to instructions. The situation became even worse when such activities started to be automated using botnets. Clickfraud is currently one of the worst methods of site promotion that causes severe negative impact on the whole Internet ecosystem. That is why search engines have to address any clickfraud attempt as quickly and effectively as possible. In order to decrease SEO's influence on search quality Yandex penalizes websites which are trying to cheat with click data and also limits the impact of click features for commercial queries ranking.

As we can see, there are several "attack vectors" which should be controlled by SE. Corresponding classifiers and restrictions are developed in Yandex to minimize effect of these attacks. Due to active counteractions from SE's side, spammers are continuously looking for new methods of cheating. Since SE developed effective algorithms of content, links, and clicks analysis, blackhats went to the areas that had not been explored well by SE - spamming dynamic content. Javascript and Actionscript allows to add anything to a webpage. HTML indexing usually does not provide SE with information about how the page will look like after JS-code interpretation. It makes possible to develop new methods of webspam which are more difficult to detect.

4. AGRESSIVE ADVERTISING

Quality of results is the most important characteristic for search engine and it has direct impact on its popularity. There are many methods of assessing the quality of search engines [3]. Quality evaluation methods based on automatic analysis of user interaction with search engine (user logs) are important for transactional queries. However, manual assessment of transactional queries is more complex, requires more time and therefore more expensive.

Same media content, files, software, etc. can be found on many different websites. Although there are multiple relevant pages that allow users to download a file or watch a video that he was looking for, it does not mean that all of them are equal for the user. For example, an entertainment web-site that provides interesting content can place advertisement to make a profit. However, its greediness leading to the extensive ad usage can make this web site very user unfriendly or unusable since required content often becomes almost inaccessible in this case. In our research of ads influence on user's experience we used "dwell time" - a well-known [2] characteristic which allows us to evaluate users' satisfaction with a webpage. We asked our search quality assessment service to assess 53200 URLs, and collected 93900 assessments. There were 4 types of assessments: ad free page, page with normal ads, "impossible to use" page, and spam (102 spam URLs were removed, 2107 pages with code 404 were removed as well). 370 websites out of 16000 (2.3%) contained pages which had been assessed as "impossible to use". Results of dwell time measurements are shown on the Figure 4.

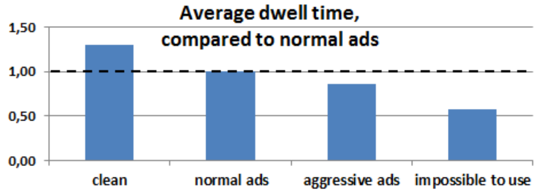


Figure 4: Average dwell time depends on the volume of ads on a web-site

It can be seen that web sites with unreasonable amount of advertisement can alienate users and provoke their complaints. The average dwell time for littered sites is 1.7 times less than for clean ones. Important to notice that this difference in time also depends on site popularity. The difference between average dwell time for normal and littered websites decreases with the growth of popularity. It brings us to more careful actions in relation to famous and popular websites if we do not want to decrease users' satisfaction.

There is an algorithm implemented at Yandex to detect such aggressive ads as well as an algorithm of average dwell time optimization. The algorithm chooses a web page without aggressive ads from several results with similar content and gives it a preference. Its use results in higher user satisfaction for transactional query results. Such strategy forces web-master to choose either to get traffic from the search engine or continue using ads excessively. This makes approach "lets create many cheap but optimized web sites with pirated content and very aggressive advertisement" (like doorways) not working. Thus, the number of aggressively monetized websites significantly decreased despite some attempts performed by their web-masters to hide ads from search engines in HTML with obfuscation and other tricks (see Figure 5).

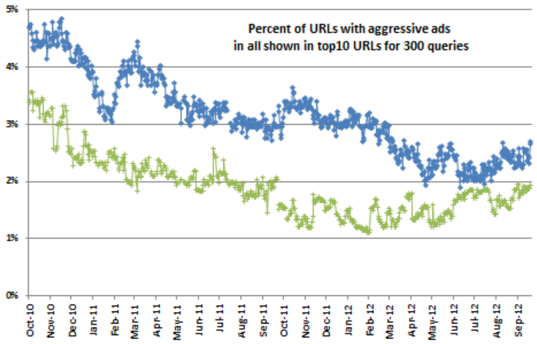


Figure 5: Number of sites with aggressive ads in russian part of web decreased in 2 times in 2 years

As we can see SE are able to detect aggressive ads in dynamic content. This ability increases blackhats' risk to loose traffic from SE. Thus spammers are faced with optimization problem: how to increase the effectiveness of a banner? Using social engineering and malware became the next step in web spam evolution.

5. WEB SPAM, FAKES, PHARMING

Micro-payment services (SMS payments) is the primary foundation for the underground economy of the Russian Internet. For example, landing pages with a paid mobile subscription is one of the most profitable way to make money through black SEO. [10]. Blackhats create very confusing scheme when it is not clear who is in charge for cashout: the mobile operator has several large partners (content and service providers), its partners have their own partners and so on. No matter how it is organized in particular, it is extremely profitable for blackhats if a user accepts a subscription or sends a paid SMS. In most cases, blackhat traffic goes to subscription forms with unclear or unreadable agreement text intended to confuse users. Social engineering techniques are widely used to increase subscription rates. Inexperienced users trust more to familiar websites and brands, and blackhats exploit their naivety very well. For example, some fraud web sites include JavaScript code that displays a fake notification message pretended to come from a popular social network, a portal or other well-known web site.

A click on fake notifications leads to special landing pages with subscription forms. Such a form is often shown as a popup banner on top of reliable web site page in the background. For example, a fake notification may contain some made-up information about Yandex lottery and promise prizes, after which user sees a mobile subscription form on the background of Yandex mainpage. Phishing tricks are being used as well. All content is being loaded dynamically using JavaScript code which is usually obfuscated. Another type of landing pages might use malicious content. There are two main approaches to attack a user in the Web: drive-by-download or social engineering. The first way exploits vulnerabilities in the user's browser, plugins, or addons. In this case the infection usually occurs unnoticed. The second approach is based on user fears, greediness, lack of experience or knowledge. It can be suggestions to download some popular software for free or "security warning" about necessary browser update (this approach has some similarity to fake antivirus [12] software). One way or another, blackhats get the opportunity to change user's system settings, modify files, etc. Currently we are observing the second wave of pharming, when *hosts* file is being modified to spoof IP addresses of popular websites. We should notice that, apparently, the most profitable *hosts* substitution leads to landing pages of the first type which are monetized via paid SMS or subscriptions. Percent of successful attacks here is higher because visually nothing changes for user. A popular website's name is being resolved into a bogus IP address, but the address bar in the browser remains the same.

6. CONCLUSIONS AND FUTURE WORK

Modern SEO and web spam techniques became more sophisticated and technologically advanced. It requires immediate actions from search engines' side. Spam detection systems should be modified with consideration of wide Javascript usage. Additional resources are needed to interpret JS-code and find corresponding features. Web spam has a tendency to converge with fraud since it demonstrates extensive use of the hacked sites, malware, botnets, phishing, and pharming. Social engineering techniques are also widely used which compels SEs to pay attention to educational projects and services. Nevertheless search engines have a great influence on the Web and continue to successfully resist web spam reducing the number of negative phenomena.

7. ACKNOWLEDGMENTS

Authors would like to thank Sergey Pimenov and Pavel Serdyukov - Yandex, LLC - for their valuable help.

8. REFERENCES

[1] J. Abernethy, O. Chapelle, and C. Castillo. Graph regularization methods for web spam detection. *Machine Learning*, 81(2):207–225, 2010.

[2] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. *Proceedings of the 29th annual international ACM SIGIR conference*, pages 19–26, 2006.

[3] R. Alia and M. S. Beg. An overview of web search evaluation methods. *Computers and Electrical Engineering*, 37(6):835–848, November 2011.

[4] M. Erdélyi, A. Garzó, and A. A. Benczúr. Web spam classification: a few features worth more. *Proceedings of the 2011 Joint WICOW/AIRWeb Workshop on Web Quality*, pages 27–34, 2011.

[5] A. Gulin and P. Karpovich. Greedy function optimization in learning to rank. *Available at: http://clck.ru/8adNE*, 2009.

[6] Y. Liu, M. Zhang, S. Ma, and L. Ru. User behavior oriented web spam detection. In *Proceedings of the 17th international conference on World Wide Web*, pages 1039–1040. ACM, 2008.

[7] K. Nikolaev, E. Zudina, and A. Gorshkov. Combining anchor text categorization and graph analysis for paid link detection. *Proceedings of the 2009 International Conference on the World Wide Web*, pages 1105–1106, 2009.

[8] J. Piskorski, M. Sydow, and D. Weiss. Exploring linguistic features for web spam detection: a preliminary study. *Proceedings of the 4th international AIRWEB workshop*, pages 25–28, 2008.

[9] F. Radlinski. Addressing malicious noise in clickthrough data. In *Learning to Rank for Information Retrieval Workshop at SIGIR*, 2007.

[10] S. Ragimova. Mobile bacchanalia. *Kompaniya (in russian)*, 601, see http://clck.ru/8adNI, 2010.

[11] N. Spirin and J. Han. Survey on web spam detection: principles and algorithms. *ACM SIGKDD Explorations Newsletter*, 13(2):50–64, 2012.

[12] B. Stone-gross, R. Abman, R. A. Kemmerer, C. Kruegel, D. G. Steigerwald, and G. Vigna. The underground economy of fake antivirus software. *University of California at Santa Barbara, Economics Working Paper Series*, 2011.