# Localized CAPTCHA Testing on Users and Farms

Ekaterina Gladkikh
Yandex
16, Leo Tolstoy st
Moscow, Russia
kglad@yandex-team.ru

Kirill Nikolaev
Yandex
16, Leo Tolstoy st
Moscow, Russia
kvn@yandex-team.ru

Mikhail Nikitin
Yandex
16, Leo Tolstoy st
Moscow, Russia
mellior@yandex-team.ru

## ABSTRACT

The paper describes the experience of resisting the large-scale solving of CAPTCHA through the CAPTCHA-farms and presents the results of experimenting with different types of textual CAPTCHA on the farm worker and casual user crowds. Localization of CAPTCHA led to cutting twice the absolute volume of CAPTCHA parsing, but introducing the semantics into the test complicated it to casual users and was not found promising.

## Categories and Subject Descriptors

K.6.5 [**Management of Computing and Information Systems**]: Security and Protection – *authentication, unauthorized access*.

## Keywords

Localized CAPTCHA; CAPTCHA-farms; usability; resisting the large-scale scraping

## 1. INTRODUCTION

Yandex is operating Russia's most popular search engine generating 62% of all search traffic in Russia [1]. Every day Yandex serves over 200 mln of user requests. Additionally, from 30 to 40 mln searches are made by various automated search bots, as estimated by the company's proprietary bot detection algorithms.

Having more than 60% of search traffic share in Russia, Yandex is an object of continuing interest from SEO-companies, which scrape Yandex search engine result pages (SERPs) for different kinds of analysis. Such scraping does not always come for free, but any related expenses are usually justified by the high cost of website promotion services provided by SEOs.

Yandex Search API [2] is the service for automated requests that was made available to satisfy the above-mentioned scraping requests within certain quota, but a considerable part of scraping load still falls on the SERP, which is supposed to be used only by actual users, not bots. A special machine learning (ML) mechanism detects automated requests on SERP and blocks them by showing CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) (Figure 1), that is created using a number of well-known protectability improving techniques - kerning overlap, local and global wrap, missing ink, additive noise, variety of fonts and formatting [5], which make the CAPTCHA recognition non-trivial for automated parsers. This is why CAPTCHA is now parsed mainly through the CAPTCHA farms – special services hiring a lot of people, mainly from regions of low-cost labor markets (e.g. China, India, Bangladesh, Vietnam) [6], who identify CAPTCHA symbols for money.

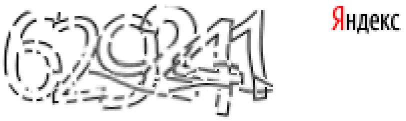

**Figure 1. Example of Yandex CAPTCHA**

As it is fairly mentioned by Motoyama et al. [6], due to the prosperity of CAPTCHA-farms, presently, CAPTCHA mechanism does not prevent large-scale automated access, but the increase of the cost of CAPTCHA solving in combination with secondary defense mechanisms (such as cookie invalidation, CAPTCHA's input limits etc.) appear to be very effective in discouraging the crowd-sourced scraping. This paper presents the Yandex experience of combating this kind of misuse.

## 2. USABILITY TESTING

As of August 2013, online experiments showed that about 80% of correct Yandex CAPTCHA inputs are the result of solving it through farms. The aim of this work was to modify CAPTCHA in such a way that it becomes more costly for CAPTCHA farms to solve it, but remains convenient to solve it for the main audience of Yandex (ex-USSR countries).

### 2.1 Experiments

Considering the origin of the majority of farm workers and the recent studies on CAPTCHA [3, 4, 7], showing that users prefer localized CAPTCHA, we have conducted a series of experiments with usability testing. A proprietary algorithm for ML-classification of users into farm-workers and casual users was used for the tests. The accuracy of classification was nearly 100% according to a verification on the labeled set of more than 3K search requests. The metrics of usability were the percentage of correct inputs by human users and the median of their input time. In total, 6 types of CAPTCHA were tested on farm workers and organic user crowd that corresponded to 3.5 mln search requests. The results of these experiments are presented in Table 1.

The Cyrillic CAPTCHAs proved to be the most problematic for the farms, as the significant share of tasks was seemingly assigned to the workers, who do not know the Cyrillic alphabet.

*WWW'14 Companion*, April 7-11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577305

Positive Cyrillic dictionary (frequently used Russian words with positive connotation like "sun", "pancakes" etc.) and Numeric types turned out to be the most convenient for organic users.

**Table 1. CAPTCHAs usability**

| Type | Users success | Users response time, s | Farm success | Farm response time, s |
|---|---|---|---|---|
| Numeric (baseline) | 96% | 16 | 63% | 12 |
| Latin random letters | 91% | 29 | 59% | 19 |
| Cyrillic random letters | 93% | 24 | 28% | 19 |
| Cyrillic dictionary | 96% | 18 | 29% | 18 |
| Positive Cyrillic dictionary | 96% | 16 | 29% | 13 |

## 3. MARKET REACTION

After introducing the Positive Cyrillic to the entire crowd, the absolute quantity of correct inputs has decreased by 80%.

The «outside» situation was evaluated by observing two stocks: antigate.com (the price of solving 1K CAPTCHAs, which depends on the number of workers online and the current demand) and xmlstock.com, where Yandex Search API's users can sell their portion of requests to those users, who need more requests than the official per user limit allows (the price for buying 1K requests).

The change of CAPTCHA type has seriously affected the work of the antigate.com. A small number of workers who know Cyrillic alphabet along with the high demand have resulted in the growth of the average price of correct solving: within the first several days the solving cost of 1k CAPTCHAs had grown by 9 times, sometimes reaching up to the unprecedented high level of 20$/1K.

From the situation on the API-limits' market it could be concluded that a part of CAPTCHA farms clients had preferred to switch to using Yandex Search API rather than accept the increased price of CAPTCHA solving. On the Figure 2 it is shown the rise of the number of xmlstock.com customers and the corresponding increase of the price for 1K requests to the Search API.

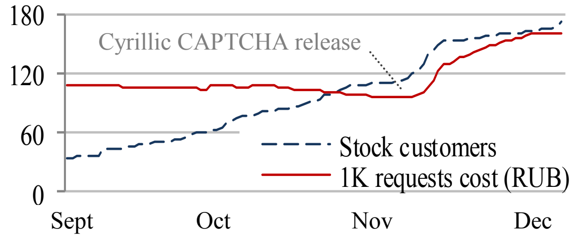

**Figure 2. Dynamics of Yandex Search API-limits' stock**

After about two weeks, the situation had stabilized, the farms had adapted to new conditions and the correct solving percentage for the farm workers had restored to approximately 60%. But considering that the solving cost had grown as well, the absolute number of correct inputs per day had decreased twice as a result of the experiment.

## 4. SEMANTIC COMPONENT IN TEST

In order to understand the nature of the restored percentage of correct inputs (whether it was due to the distribution of the tasks to new workers from the ex-USSR countries or the farms have used a virtual Cyrillic keyboard for all the workers), the entire crowd was loaded with the CAPTCHAs made of common words in Cyrillic with a missed vowel.

After the release of this type of CAPTCHA, the percentage of successful inputs decreased for the human crowd by 13% and by 13.5% for the farm-workers. From the fact that it resulted into an almost the same decrease of solving quality for the both crowds, we conclude that the nature of the restore is the consequence of the distribution of the solving tasks to Russian-speaking workers.

## 5. CONCLUSION

Localized CAPTCHA reduced the percentage of successful inputs for farm workers, but after the farms adapted, the success percentage for the farm workers had restored. Nevertheless, the absolute volumes of CAPTCHA parsing have been cut twice in comparison with the initial conditions and the familiar level of comfort for organic users was still preserved.

Introducing the semantic component to the test complicated CAPTCHA for users as well as for farms but did not disable any significant part of farm workers. Thus, the strengthening of semantics in CAPTCHA is not the promising direction in fighting against CAPTCHA-farms in modern conditions.

## 6. ACKNOWLEDGMENTS

Special thanks to Pavel Serdyukov for his careful proof-reading and numerous useful comments.

## 7. REFERENCES

[1] Yandex.Company - http://company.yandex.com/

[2] Yandex.XML - http://xml.yandex.com/

[3] Fidas, C., Voyiatzis, A.G. 2013. On Users' Preference on Localized vs. Latin-Based CAPTCHA Challenges. *Human-Computer Interaction – INTERACT* (2013), 358-365.

[4] Tangmanee, C., Sujarit-apirak, P. 2013. Attitudes towards CAPTCHA: A Survey of Thai Internet Users. *The Journal of Global Business Management Volume 9 Number 2* (June 2013), Special Edition, 29-41.

[5] Thomas, A., Punera, K., Kennedy, L., Tseng, B., Chang, Y. Framework for Evaluation of Text CAPTCHAs. *WWW 2013 Companion, May 13–17, Rio de Janeiro, Brazil*, 159-160.

[6] Motoyama, M., Levchenko, K., Kanich, C., McCoyl, D., Voelker, G.M., Savage, S. 2010. Re:CAPTCHAs – Understanding CAPTCHA-Solving Services in an Economic Context. *USENIX Security'10 Proceedings of the 19th USENIX conference on Security*

[7] Yan, J., El Ahmad, A.S. 2008. Usability of CAPTCHAs or usability issues in CAPTCHA design. *In SOUPS '08, New York, NY, USA (2008), 44-52*