# Report on Cirrus Search A/B Test

TextCat Language Detection on English Wikipedia

*Mikhail Popov (Analysis & Report)*
*Trey Jones (Engineering & Review)*
*Erik Bernhardson (Engineering)*
*David Causse (Engineering)*
*Stas Malyshev (Engineering)*
*Dan Garry (Product Management)*
*Deborah Tankersley (Product Management)*

*July 13, 2016*

## Executive Summary

Our current efforts to increase relevancy of search results rely on language detection (via the TextCat library) to provide the user with results from a potentially more relevant alternate language. For example, users who were enrolled in the test group and searched English Wikipedia in French received additional "interwiki" results from French Wikipedia.

The test groups not only had a substantially lower zero results rate (57% in control group vs 46% in the two test groups), but they had a higher clickthrough rate (44% in the control group vs 49-50% in the two test groups), indicating that we may be providing users with relevant results that they would not have gotten otherwise. Interestingly, users first clicked on the first and second search results from same-language wiki, rather than the potentially more relevant interwiki results.

We recommend continuing our work with TextCat and possibly deploying it to production. We should consider recording the confidence of the language detection, as there may be a correlation between the confidence – a proxy for the potential relevancy of the results – and the likelihood that the user will click on the interwiki result(s).

## Introduction

In our investigation of search queries, we have found that sometimes people write search queries in a language different from the language of the wiki they are searching. Sometimes this works (some articles may have translations on the page) and sometimes it does not. So we asked ourselves, "What if we could detect the language of the query and then if the user did not get many – if any – results on their current wiki, what if just searched the potentially correct wiki and gave the user THOSE results?" To that end, we decided to proceed with adding language detection to Cirrus search (see T118278).

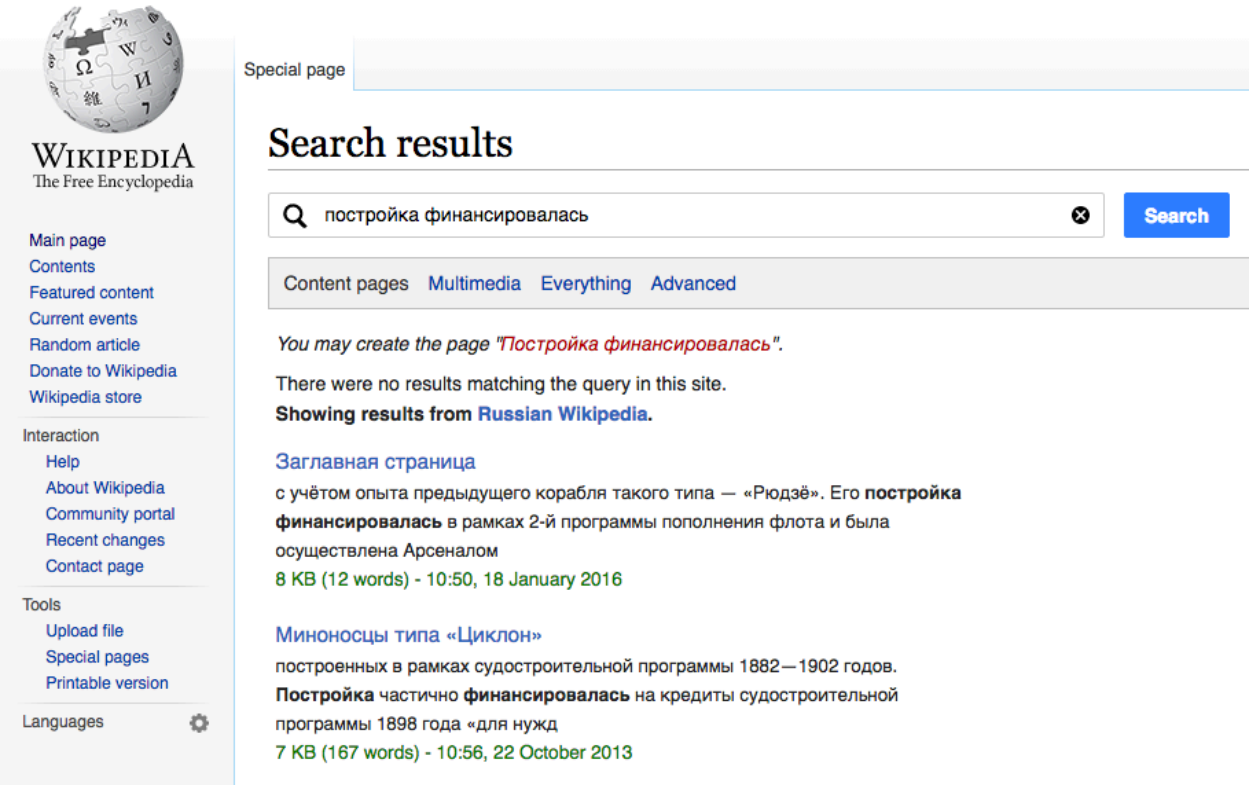


*Figure 1: A screenshot showing what happens when a user searches English Wikipedia using a query written in Russian.*

TextCat is a software library for detecting language based on n-gram text categorization, which we ported to PHP for use in Cirrus search. In this initial A/B test of the TextCat software, we deployed the code to English, French, Spanish, Italian, and German Wikipedias ("enwiki", ..., "dewiki" = "same-wiki") to determine how people engage with the "interwiki" results (results from a Wikipedia in another language).

## Methods

Test group users who received less than 3 same-wiki results and whose language we were able to detect (using either TextCat or combination of TextCat and Accept-Language header) received interwiki results. We still tried to detect a language for users in the control group so they experience the same lag that the users in the test groups did; this also makes for more valid comparisons. Our initial attempt ran from 16 March 2016 to 31 May 2016, but a few issues with the data (see T137158#2359494) meant that we had to resolve them and then restart the test. The second attempt ran from 16 June 2016 to 05 July 2016. A total of 49,623 sessions (users) and 116,462 searches were recorded using the MediaWiki Event Logging system. The data analysis was performed using R in RStudio and the packages: magrittr, uaparser, dplyr, tidyr, and BCDA.

## Results

In Table 1, we see that the proportion of sessions (which perform multiple searches) for which we could detect a language is nearly double the proportion of sessions for which we could not. Nearly the opposite holds for proportion of individual searches.

*Table 1: Number of sessions and individual searches within them where we detected or did not detect a language.*

| wiki | Detected a language | searches | searches (proportion) | sessions | sessions (proportion) |
|---|---|---|---|---|---|
| dewiki | Detected | 2150 | 0.402 | 1397 | 0.673 |
| dewiki | Did not detect | 3198 | 0.598 | 678 | 0.327 |
| enwiki | Detected | 5437 | 0.388 | 3562 | 0.650 |
| enwiki | Did not detect | 8593 | 0.612 | 1922 | 0.350 |
| eswiki | Detected | 1113 | 0.381 | 681 | 0.634 |
| eswiki | Did not detect | 1805 | 0.619 | 393 | 0.366 |
| frwiki | Detected | 632 | 0.479 | 422 | 0.676 |
| frwiki | Did not detect | 688 | 0.521 | 202 | 0.324 |
| itwiki | Detected | 457 | 0.463 | 266 | 0.662 |
| itwiki | Did not detect | 531 | 0.537 | 136 | 0.338 |

While the zero results rate is not the key metric of interest to us in this particular test, we still want to make sure that we are seeing the numbers we might expect. In Table 2, we can see that groups who received the additional interwiki results had a substantially lower zero results rate, which makes sense.

*Table 2: Searches where we could detect a language had a much lower zero results rate when using interwiki results via TextCat and Accept-Language header language detection.*

| group | zero results | some results |
|---|---|---|
| a (control) | 3237 (57.05%) | 2437 (42.95%) |
| b (textcat) | 2561 (45.57%) | 3059 (54.43%) |
| c (textcat + accept-lang) | 3239 (46.33%) | 3752 (53.67%) |

In the clickthrough analyses that follow, we only use sessions for which we detected a language and which had some search results.

*Table 3: Number of sessions (for which we detected a language in one of the search queries) per group per wiki.*

| wiki | a (control) | b (textcat) | c (textcat + accept-lang) |
|---|---|---|---|
| dewiki | 261 (18.68%) | 488 (34.93%) | 648 (46.39%) |
| enwiki | 762 (21.39%) | 1341 (37.65%) | 1459 (40.96%) |
| eswiki | 141 (20.70%) | 252 (37.00%) | 288 (42.29%) |
| frwiki | 66 (15.64%) | 150 (35.55%) | 206 (48.82%) |
| itwiki | 33 (12.41%) | 103 (38.72%) | 130 (48.87%) |

*Table 4: Proportion of sessions that clicked specifically on a 'same-language' (En/Fr/Es/It/De) Wikipedia result. The two test groups ('b' and 'c') have been combined into a single group.*

| | dewiki | enwiki | eswiki | frwiki | itwiki | overall |
|---|---|---|---|---|---|---|
| controls | 0.479 | 0.449 | 0.411 | 0.333 | 0.424 | 0.444 |
| test | 0.459 | 0.436 | 0.346 | 0.371 | 0.425 | 0.427 |

*Table 5: Clickthrough rates across the five Wikipedias. The two test groups ('b' and 'c') have been combined into a single 'test' group.*

| | dewiki | enwiki | eswiki | frwiki | itwiki | overall |
|---|---|---|---|---|---|---|
| controls | 0.479 | 0.449 | 0.411 | 0.333 | 0.424 | 0.444 |
| test | 0.523 | 0.502 | 0.394 | 0.407 | 0.506 | 0.489 |

*Table 6: Clickthrough rates across the five Wikipedias. For the groups 'b' and 'c', these clickthrough rates account for both types of results that could be clicked – same-wiki and interwiki.*

| | dewiki | enwiki | eswiki | frwiki | itwiki | overall |
|---|---|---|---|---|---|---|
| a (control) | 0.479 | 0.449 | 0.411 | 0.333 | 0.424 | 0.444 |
| b (textcat) | 0.512 | 0.512 | 0.337 | 0.413 | 0.515 | 0.487 |
| c (textcat + accept-lang) | 0.531 | 0.493 | 0.444 | 0.403 | 0.500 | 0.490 |

*Table 7: Counts and proportions of searches with some or no En/Fr/Es/It/De Wikipedia results and some or no interwiki results where we detected a language and the user clicked on a result.*

| | some same-wiki results | zero same-wiki results |
|---|---|---|
| some interwiki results | 620 (37.64%) | 425 (25.80%) |
| zero interwiki results | 602 (36.55%) | 0 (0.00%) |

## Conclusion

This initial evidence suggests that using TextCat to detect language and present the users with additional results from the Wikipedia in the detected language has a benefit for the users. We suggest recording the confidence of the language detection. That is, perhaps the number of searches where the user clicked on an interwiki result first rather than current wiki is high for searches where we have a very high confidence ("Oh yeah, that's definitely the language they are searching in but just are on the wrong wiki.") of correctly detecting the user's language, compared to those searches where our detection can be best described as, "Well, I guess this could be the language they were trying to search in." Trey Jones began working on this.

*Table 8: The results of a Bayesian analysis using the Beta-Binomial model of clickthrough rates by group. The test group(s) were more likely to clickthrough on a result than the controls (who did not receive interwiki results).*

| **Outcome** | $N_1$ | $N_2$ | $P_1$ | $P_2$ | **Difference** $(P_1 - P_2)$ | **Relative Risk** | **Odds Ratio** |
|---|---|---|---|---|---|---|---|
| Clicked on a same-wiki result | 5065 | 1263 | 42.67% (41.30%, 44.04%) | 44.43% (41.73%, 47.17%) | -1.77% (-4.86%, 1.28%) | 0.96 (0.90, 1.03) | 0.93 (0.82, 1.05) |
| Clicked on a result ('test' vs 'controls') | 5065 | 1263 | 48.88% (47.49%, 50.24%) | 44.41% (41.63%, 47.21%) | 4.47% (1.36%, 7.55%) | 1.10 (1.03, 1.18) | 1.20 (1.06, 1.36) |
| Clicked on a result ('b' vs 'a') | 2334 | 1263 | 48.71% (46.70%, 50.65%) | 44.43% (41.66%, 47.19%) | 4.28% (0.91%, 7.61%) | 1.10 (1.02, 1.18) | 1.19 (1.04, 1.36) |
| Clicked on a result ('c' vs 'a') | 2731 | 1263 | 49.02% (47.14%, 50.90%) | 44.42% (41.68%, 47.14%) | 4.61% (1.29%, 7.87%) | 1.10 (1.03, 1.19) | 1.21 (1.05, 1.37) |
| Clicked on a result ('b' vs 'c') | 2334 | 2731 | 48.73% (46.70%, 50.72%) | 49.02% (47.11%, 50.88%) | -0.28% (-3.04%, 2.43%) | 0.99 (0.94, 1.05) | 0.99 (0.89, 1.10) |