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 11, 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. That is, users who were enrolled in the test group and searched English Wikipedia ("enwiki") in French received additional "interwiki" results from French Wikipedia.

The test groups not only had a substantially lower zero results rate (58% in control group vs 43-45% 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 enwiki, rather than the potentially more relevant interwiki results.

We recommendation is to proceed with the test on other Wikipedias, but to also record the confidence of the language detection. We suspect 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).

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 Wikipedia ("enwiki") to determine how people engage with the "interwiki" results (results from a Wikipedia in another language).

Methods

Users who received less than 3 enwiki results and whose language we were able to detect (using either TextCat or combination of TextCat and Accept-Language header) received interwiki results. 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.

Detected a language	searches	searches (proportion)	sessions	sessions (proportion)
Detected	3897	0.397	2600	0.656
Did not detect	5914	0.603	1366	0.344

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)	1264 (57.77%)	924 (42.23%)
b (textcat)	1022 (44.55%)	1272 (55.45%)
c (textcat + accept-lang)	1108 (43.16%)	1459 (56.84%)

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.

Group	Sessions
a (control)	531
b (textcat)	993
c (textcat + accept-lang)	1076

Table 4: Proportion of sessions that clicked specifically on an English Wikipedia result vs proportion that did not click on any result. The two test groups ('b' and 'c') have been combined into a single group.

	Clicked an enwiki result	Did not
controls	0.441	0.559
test	0.449	0.551

Table 5: Proportions of clicks on results (enwiki and interwiki). The two test groups ('b' and 'c') have been combined into a single group.

	Clicked a result	Did not
controls	0.441	0.559
test	0.496	0.504

Table 6: Proportion of sessions that clicked on a result vs that did not click on any result. For the groups 'b' and 'c', these clickthrough rates account for both types of results that could be clicked – enwiki and interwiki.

	Clicked a result	Did not
a (control)	0.441	0.559
b (textcat)	0.497	0.503
c (textcat + accept-lang)	0.495	0.505

Table 7: Counts and proportions of searches with some or no enwiki and interwiki results where we detected a language and the user clicked on a result.

	some enwiki results	zero enwiki results
some interwiki results	184 (28.13%)	196 (29.97%)
zero interwiki results	274 (41.90%)	0 (0.00%)

An interesting thing to note is that of the 654 searches made by the two test groups 'b' and 'c' that had (1) had a language detected, and (2) a clickthrough: only 28.13% (184) searches had both interwiki and enwiki results. We were curious how those users engaged with their mixed results, so we looked into and found that users almost overwhelmingly initially clicked on the 1-2 enwiki results that show up first, rather than the interwiki results shown underneath the enwiki ones. This is not a particularly new finding, as we have known and seen for a while that a vast majority of users just click on the first couple results. This is interesting because it introduces two interesting possibilities: either the interwiki results are not as relevant as we hope them to be, or users just click on the first or second result even if the results underneath are actually, objectively more relevant because they were fetched from the Wikipedia of the query's language.

Table 8: Do people just click on the first link even if the better result is the interwiki result that's lower in the list? Seems like it.

enwiki results	first clicked on	1st result	2nd result	3rd result	4th result or lower
1	an enwiki result	72 (39.13%)	5 (2.72%)	1 (0.54%)	1 (0.54%)
1	an interwiki result	26 (14.13%)	1 (0.54%)	2 (1.09%)	4 (2.17%)
2	an enwiki result	47 (25.54%)	6 (3.26%)	0 (0.00%)	1 (0.54%)
2	an interwiki result	12 (6.52%)	4 (2.17%)	0 (0.00%)	2 (1.09%)

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 proceeding with a follow-up test, but to also record 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."

Table 9: 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 an enwiki result	2069	531	44.90% (42.74%, 47.06%)	44.10% (39.93%, 48.31%)	0.81% (-3.95%, 5.45%)	1.02 (0.92, 1.14)	1.04 (0.85, 1.25)
Clicked on a result ('test' vs 'controls')	2069	531	49.63% (47.46%, 51.76%)	44.06% (39.78%, 48.38%)	5.57% (0.72%, 10.30%)	1.13 (1.01, 1.26)	1.26 (1.03, 1.52)
Clicked on a result ('b' vs 'a')	993	531	49.74% (46.66%, 52.71%)	44.09% (39.83%, 48.31%)	5.65% (0.46%, 10.90%)	1.13 (1.01, 1.27)	1.26 (1.02, 1.55)
Clicked on a result ('c' vs 'a')	1076	531	49.53% (46.53%, 52.52%)	44.07% (39.87%, 48.28%)	5.46% (0.21%, 10.58%)	1.13 (1.00, 1.26)	1.25 (1.01, 1.53)
Clicked on a result ('b' vs 'c')	993	1076	49.78% (46.66%, 52.82%)	49.51% (46.47%, 52.48%)	0.26% (-4.01%, 4.60%)	1.01 (0.92, 1.10)	1.01 (0.85, 1.20)