Second BM25 A/B Test Analysis

Chelsy Xie (Analysis & Report)
Erik Bernhardson (Engineering)
David Causse (Engineering & Review)
Trey Jones (Engineering & Review)
Mikhail Popov (Review)
Deb Tankersley (Product Management)

06 January 2017

Executive Summary

In order to assess the efficacy of BM25 in space-less language, Discovery's Search team has decided to conduct a second A/B test in Chinese, Japanese and Thai Wikipedias. We observed that the test group that used per-field query builder with incoming links and pageviews as query-independent factors had a much better Zero Result Rate but slightly worse PaulScores, large decrease in clickthrough rate, and fewer users clicked on the first result first, which indicates that we are showing test group users worse results. However, longer dwell-time and fewer query reformulations show that test group users might actually like the results they are getting better than the control group in that respect. We recommend deploying BM25 for all wikis but not reindexing projects in space-less languages for now.

Background

To improve the relevancy of search results, Discovery's Search team decided to try a new document-ranking function called Okapi BM25 (BM stands for Best Matching), and ran an A/B test from August 30 to September 10 to assess the efficacy of the proposed switch. The analysis showed that BM25 ranking with incoming links and pageviews as query-independent factors appears to give users results that are more relevant and that they engage with more.

However, we then realized that our analysis chain is sub-optimal for space-less language queries, which will break words on every characters for the plain field. Therefore, we ran a second A/B test for Chinese, Japanese and Thai Wikipedias to test whether the new per-field BM25 builder is sufficiently worse with those languages. We are primarily interested in:

- **Zero results rate**, the proportion of searches that yielded zero results (smaller is usually better)
- **Users' engagement** with the search results, measured as the clickthrough rate (bigger is better)
- **PaulScore**, a metric of search results' relevancy that relies on the position of the clicked result[s] (bigger is better); see PaulScore Definition for more details

- **Query reformulation** – one way to think about the strength of our search engine is how many times the user reformulates their query; if a user in the test group has to reformulate their query many more times to get the results they are interested in, then maybe the change is for the worse
- **Dwell Time**, the time (seconds) that users stayed on the pages they visited by clicking on the search results (bigger is better)
- **Scroll** – if users scroll on the visited page, they are more likely to engage with the contents

Data

For Chinese Wikipedia (zhwiki) and Japanese Wikipedia (jawiki), users had a 1 in 16 chance of being selected for anonymous tracking according to our TestSearchSatisfaction2 #15922352 event logging schema. Those users who were randomly selected to have their sessions anonymously tracked then had a 12 in 13 chance of being selected for the BM25 test. For Thai Wikipedia (thwiki), users had a 1 in 5 chance of being selected for search satisfaction tracking and then had a 38 in 39 chance of being selected for the BM25 test due to fewer visitors to that particular project. The sampled sessions were evenly put into a control group (tf-idf) and a test group (using per-field query builder with incoming links and pageviews as QIFs); see the first BM25 test report for more details. The test was deployed on October 27th and ran for a week for zhwiki and jawiki, but until November 15th for thwiki specifically.

The full-text (as opposed to auto-complete) searching event logging data was extracted from the database using this script. We collected a total of 230.2K events from 36.1K unique sessions. See Table 1 for counts broken down by wiki and test group.

wiki	Test group	Search sessions	Events recorded
jawiki	Control Group (tf–idf)	7,579	54,261
jawiki	Using per-field query builder with incoming links and pageviews as QIFs	7,610	59,808
thwiki	Control Group (tf–idf)	3,889	18,954
thwiki	Using per-field query builder with incoming links and pageviews as QIFs	4,055	21,217
zhwiki	Control Group (tf–idf)	6,510	37,561
zhwiki	Using per-field query builder with incoming links and pageviews as QIFs	6,478	38,382
All Wikis	Total	36,121	230,183

Table 1: Counts of sessions anonymously tracked and events collected during the second A/B test (Oct 27 - Nov 15).

An issue we noticed with the event logging is that when the user goes to the next page of search results or clicks the Back button after visiting a search result, a new page ID is generated for the search results page. The page ID is how we connect click events to search result page events. There is currently a Phabricator ticket (T146337) for addressing these issues. For this analysis, we de-duplicated by connecting search engine results page (searchResultPage) events that have the exact same search query, and then connected click events together based on the searchResultPage connectivity.

After de-duplicating, we collapsed 213.4K (searchResultPage and click) events into 64.5K searches. See Table 2 for counts broken down by wiki and test group.

wiki	Test group	Search sessions	Searches recorded	Events recorded
jawiki	Control Group (tf–idf)	7,579	13,839	50,845
jawiki	Using per-field query builder with incoming links and pageviews as QIFs	7,610	13,982	56,097
thwiki	Control Group (tf–idf)	3,889	7,005	16,379
thwiki	Using per-field query builder with incoming links and pageviews as QIFs	4,055	7,086	18,481
zhwiki	Control Group (tf–idf)	6,510	11,450	35,262
zhwiki	Using per-field query builder with incoming links and pageviews as QIFs	6,478	11,173	36,373
All Wikis	Total	36,121	64,535	213,437

Table 2: After searchResultPage De-duplication, Counts of sessions anonymously tracked and events collected during the second A/B test (Oct 27 - Nov 15).

There are 22.1K visitPage events (When the user clicks a link in the results a visitPage event is created). See Table 3 for counts broken down by wiki and test group.

wiki	Test group	Visited pages
jawiki	Control Group (tf–idf)	5,431
jawiki	Using per-field query builder with incoming links and pageviews as QIFs	5,954
thwiki	Control Group (tf–idf)	1,554
thwiki	Using per-field query builder with incoming links and pageviews as QIFs	1,776
zhwiki	Control Group (tf–idf)	3,591
zhwiki	Using per-field query builder with incoming links and pageviews as QIFs	3,807
All Wikis	Total	22,113

Table 3: Counts of visited pages from search sessions anonymously tracked during the second A/B test (Oct 27 - Nov 15).

Results

Zero Results Rate

In Figure 1, we see that the test group that used BM25 with incoming links and pageviews as query-independent factors had a significantly lower zero results rate (ZRR). Smaller ZRR is usually better, but looking at the PaulScore, engagement and First Clicked Result's Position, we doubt that it came at the cost of relevance and engagement with the results. Figure 2 shows that zhwiki had the largest ZRR difference between control and test group.

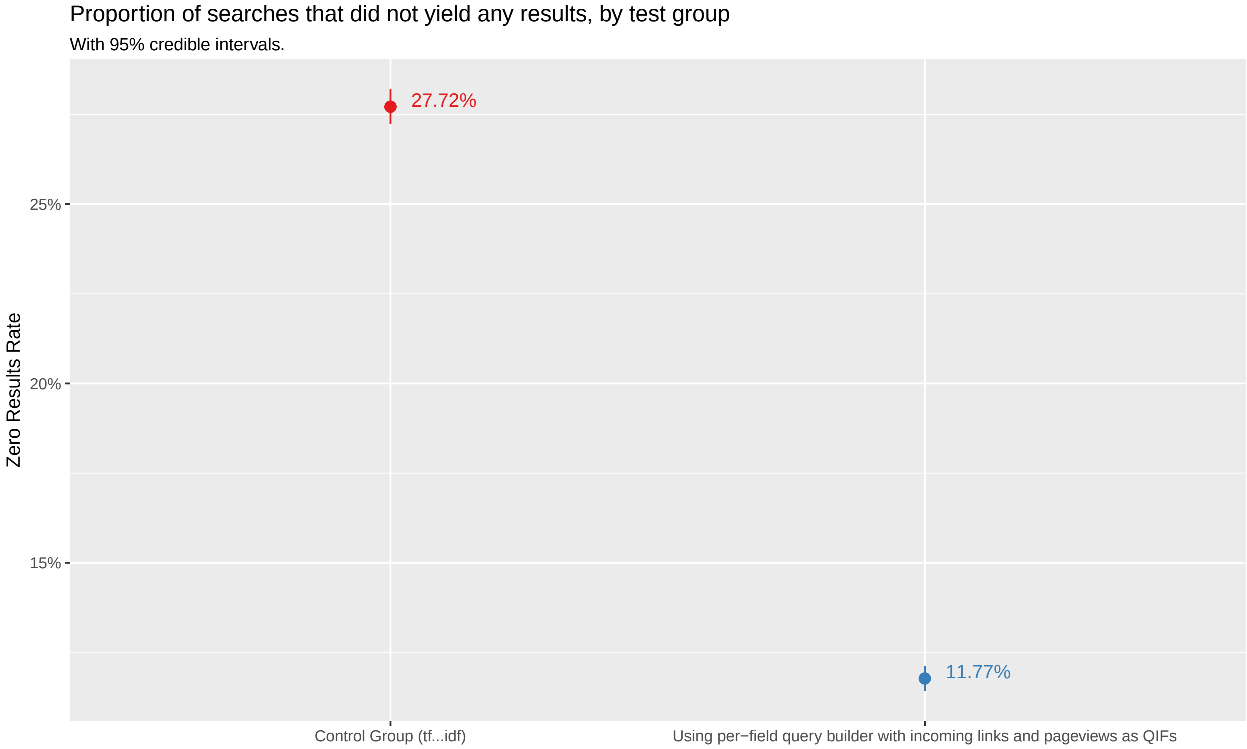

Figure 1: Zero results rate is the proportion of searches in which the user received zero results. Broken down by test group.

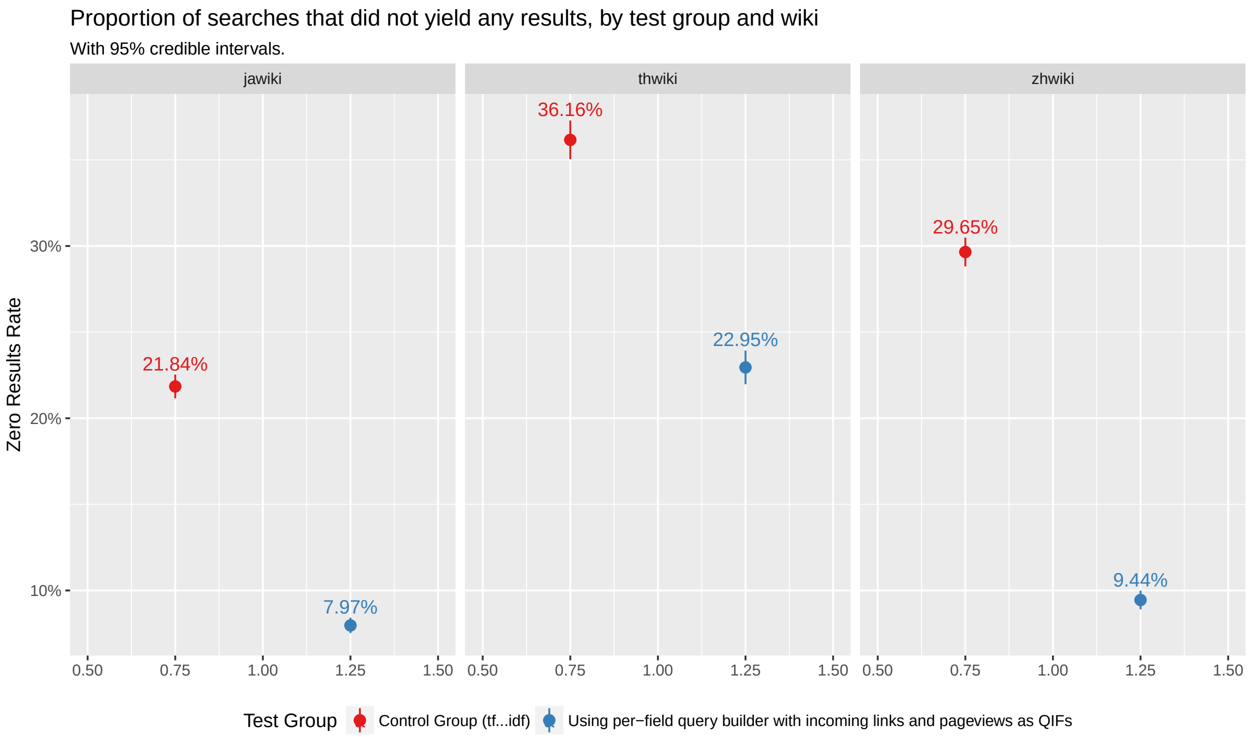

Figure 2: Zero results rate is the proportion of searches in which the user received zero results. Broken down by test group and wiki.

PaulScore

In Figure 3, we see that the test group had slightly lower PaulScores, which indicates that the results were less relevant. The difference is not significant when F = 0.9. This make sense because the smaller the value of scoring factor, the more weight put on the first few results, with lower ranking results counting for almost nothing. F=0.9 takes into account the broadest range of result rankings, and thus is less likely to change as dramatically. Figure 4 shows that zhwiki had the largest PaulScore differences between control and test group.

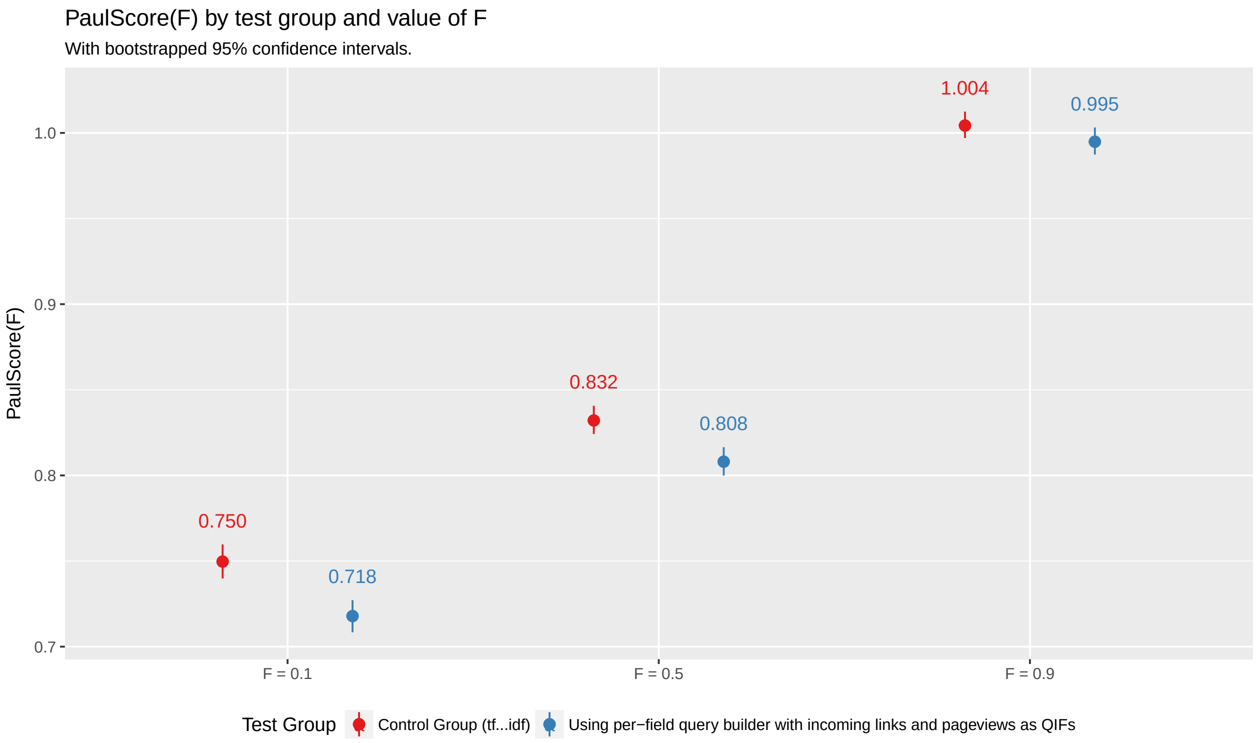

Figure 3: Average per-group PaulScore for various values of F (0.1, 0.5, and 0.9) with bootstrapped confidence intervals.

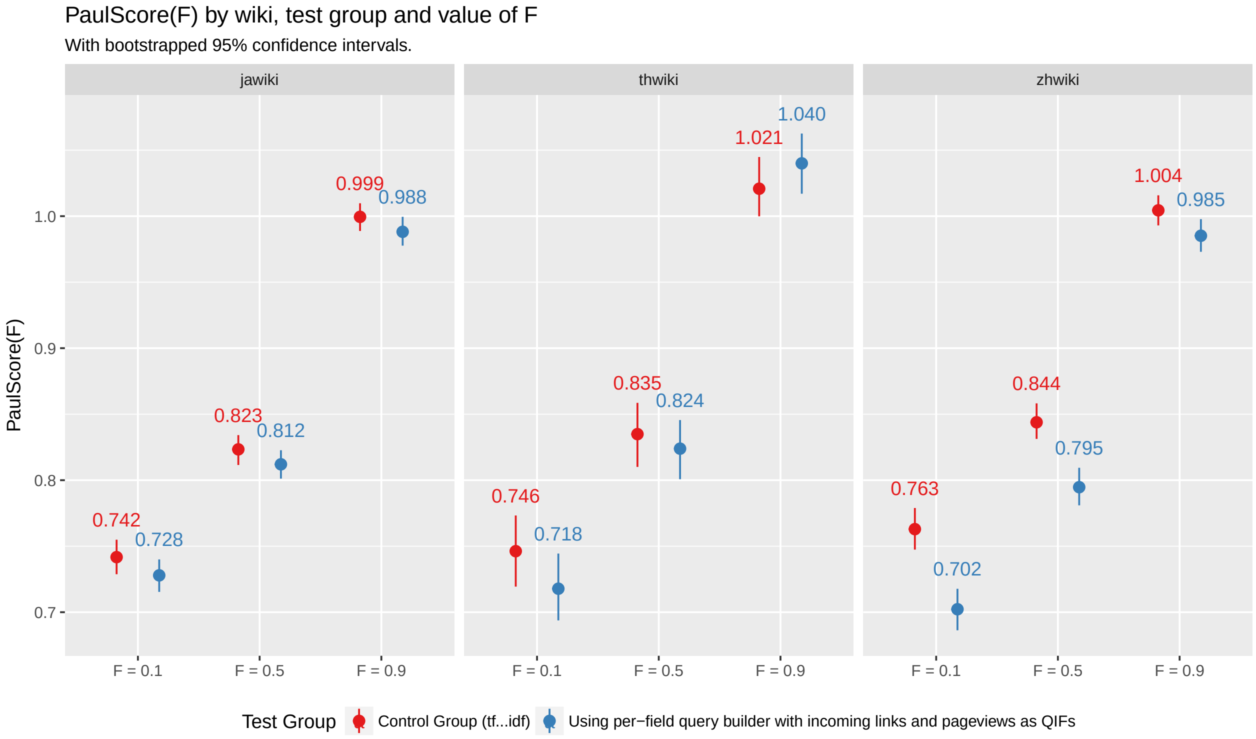

Figure 4: Average per-group PaulScore for various values of F (0.1, 0.5, and 0.9) with bootstrapped confidence intervals. Broken down by test group and wiki.

Engagement

In Figure 5, we see that the test group had a significantly lower clickthrough rate, which means users are less engaged with their search results. Again, zhwiki shows the largest discrepancy between control and test group in Figure 6.

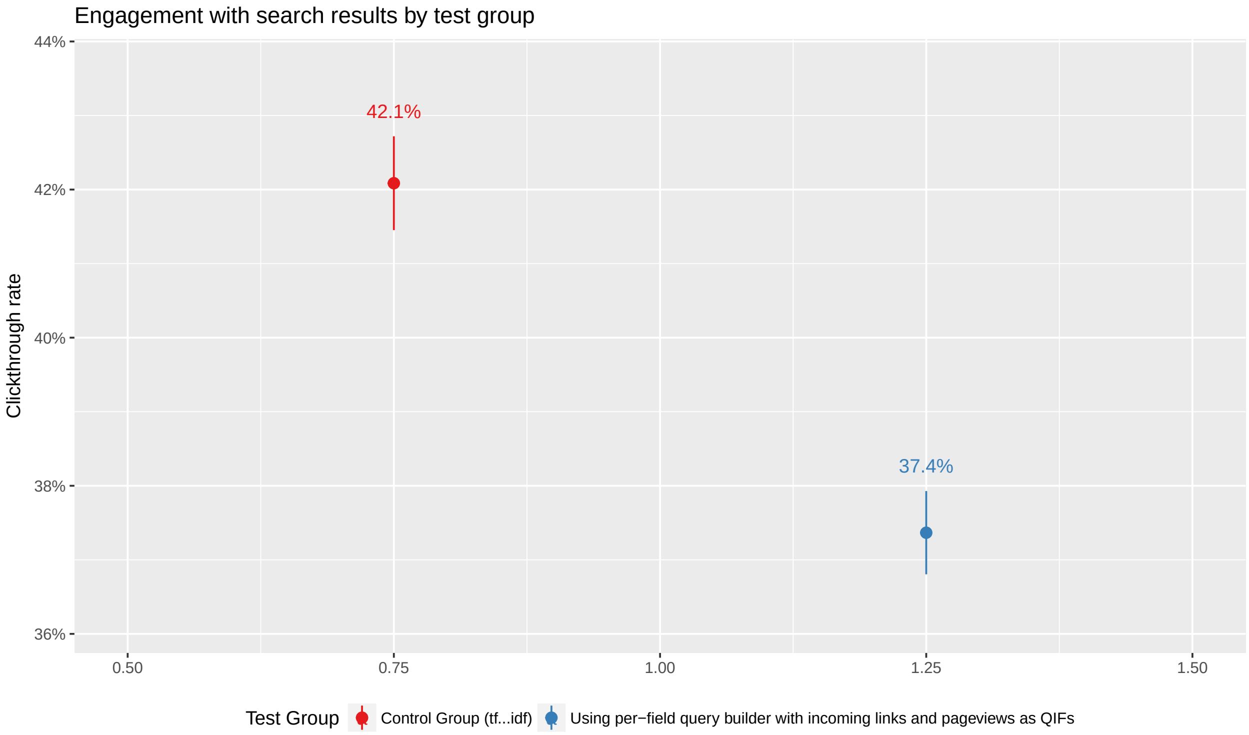

Figure 5: Clickthrough rates of test groups.

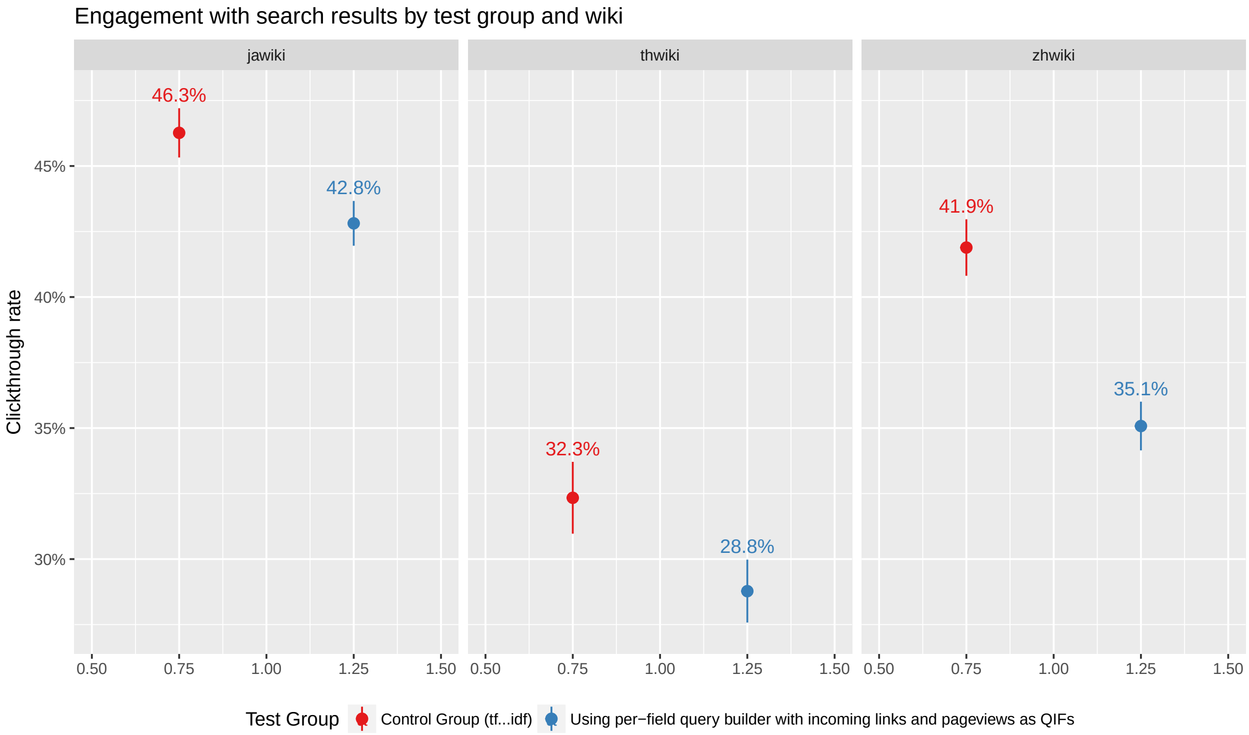

Figure 6: Clickthrough rates broken down by test group and wiki.

First Clicked Result's Position

In Figure 7, we see that test group users were less likely to click on the first search result first than the control group. Figure 8 shows that only zhwiki users first clicked on the first result at a significantly lower rate, which indicates that the results were less relevant.

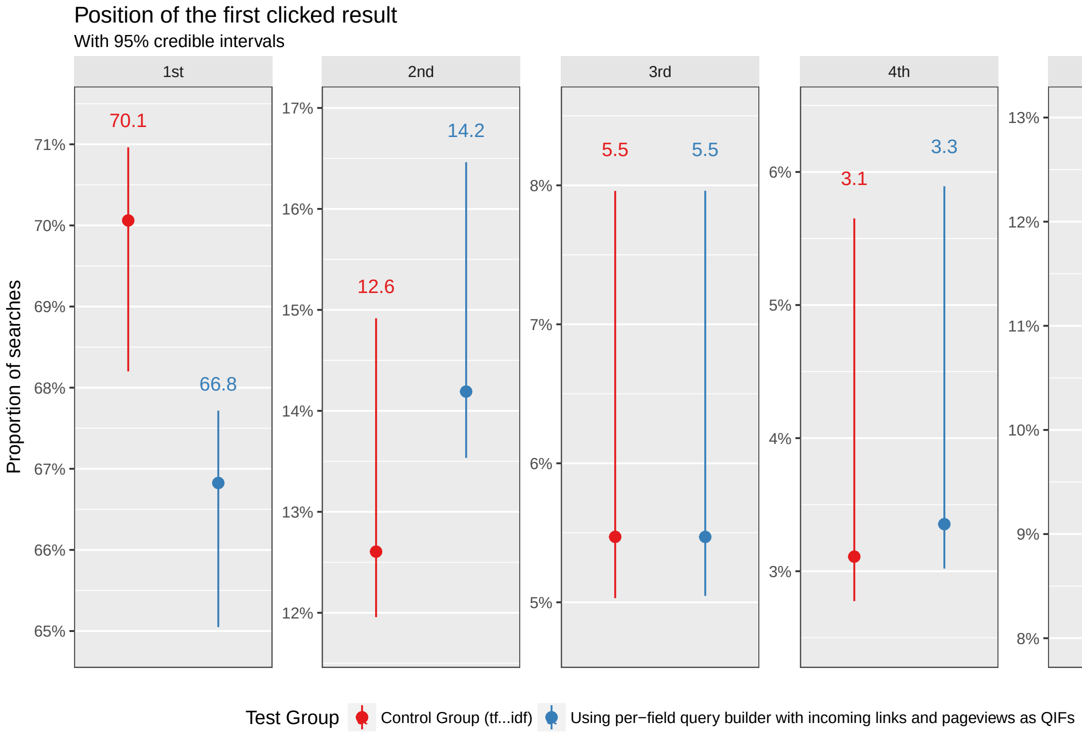

Figure 7: First clicked result's position by test group.

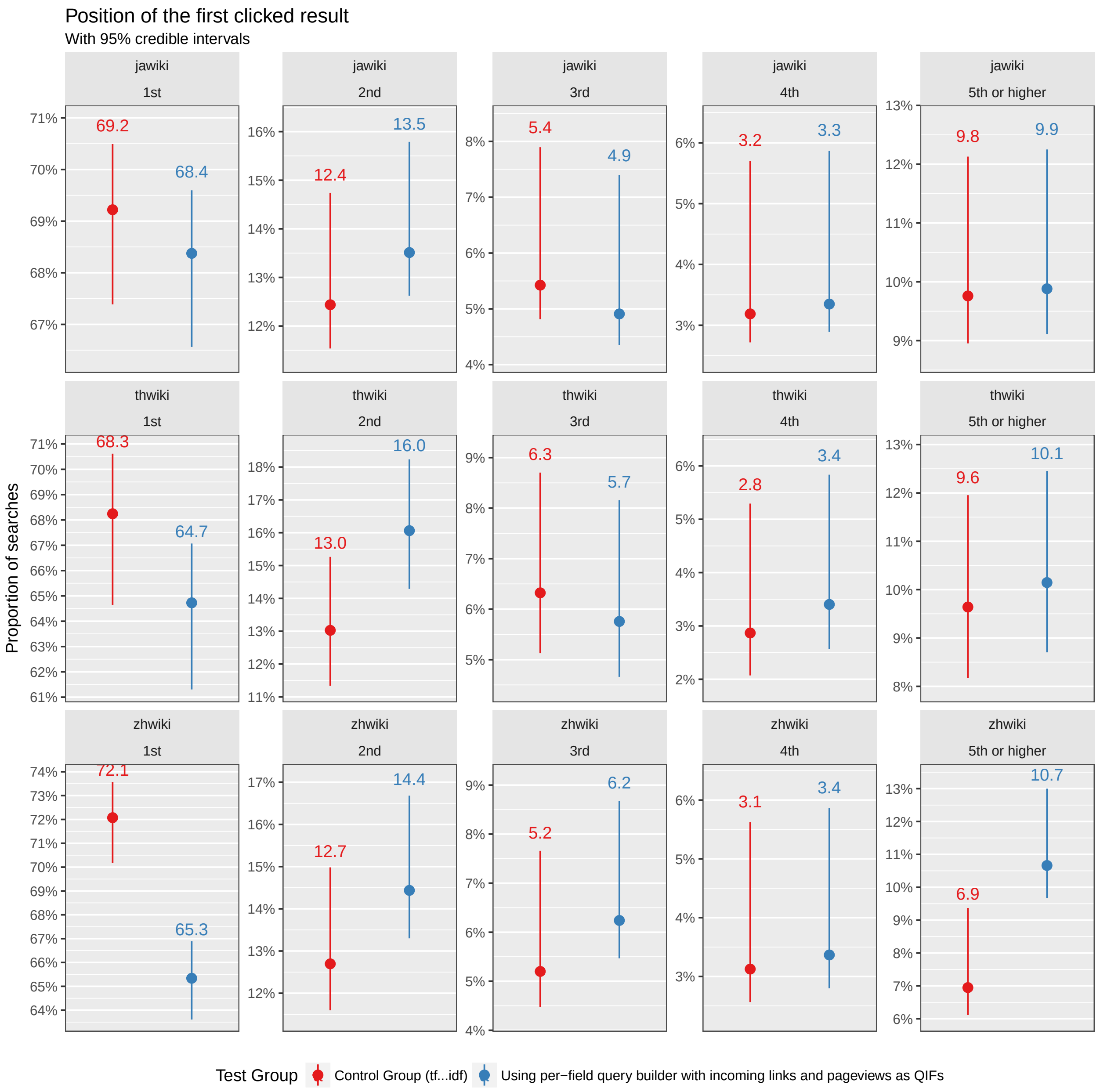

Figure 8: First clicked result's position by test group and wiki.

Dwell Time per Visited Page

Figures 9 and 10 show the survival curve for each test group and wiki. Except zhwiki, users are more likely to stay longer on visited pages, which implies the results in test group are more relevant for jawiki and thwiki.

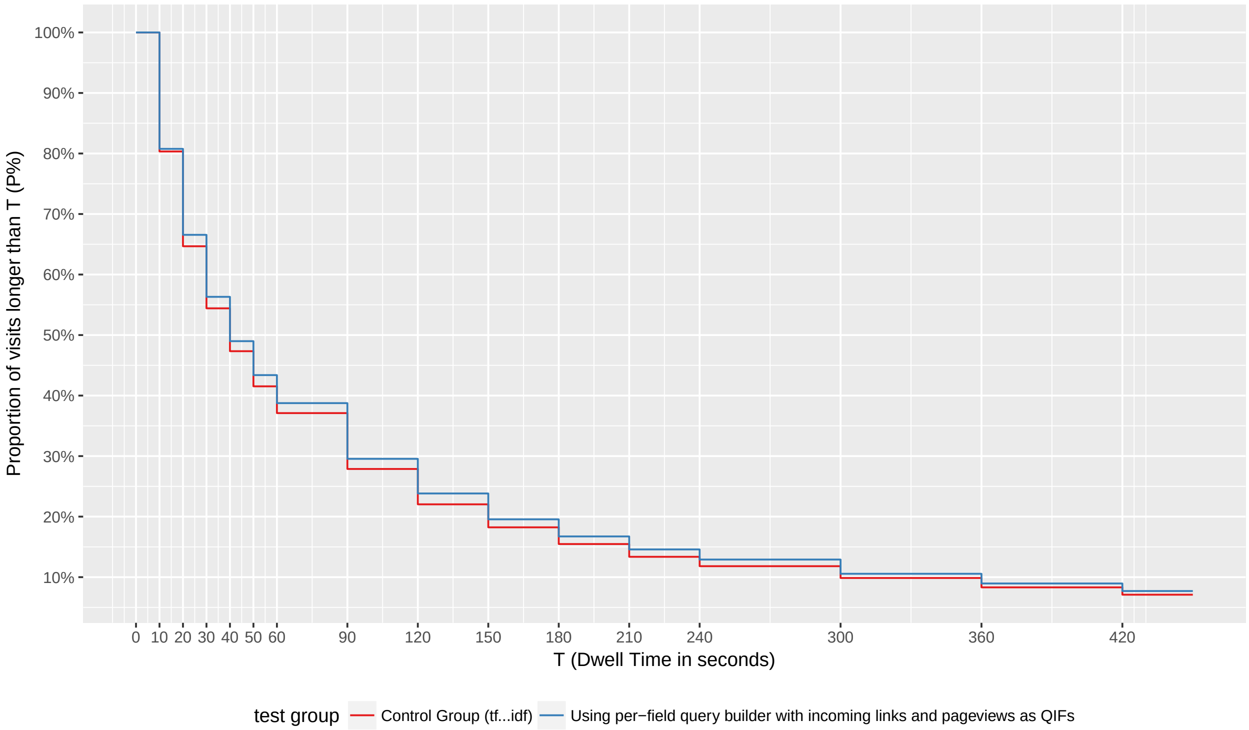

Figure 9: At time T, at most P% of users still stay on their visited pages. Broken down by test group.

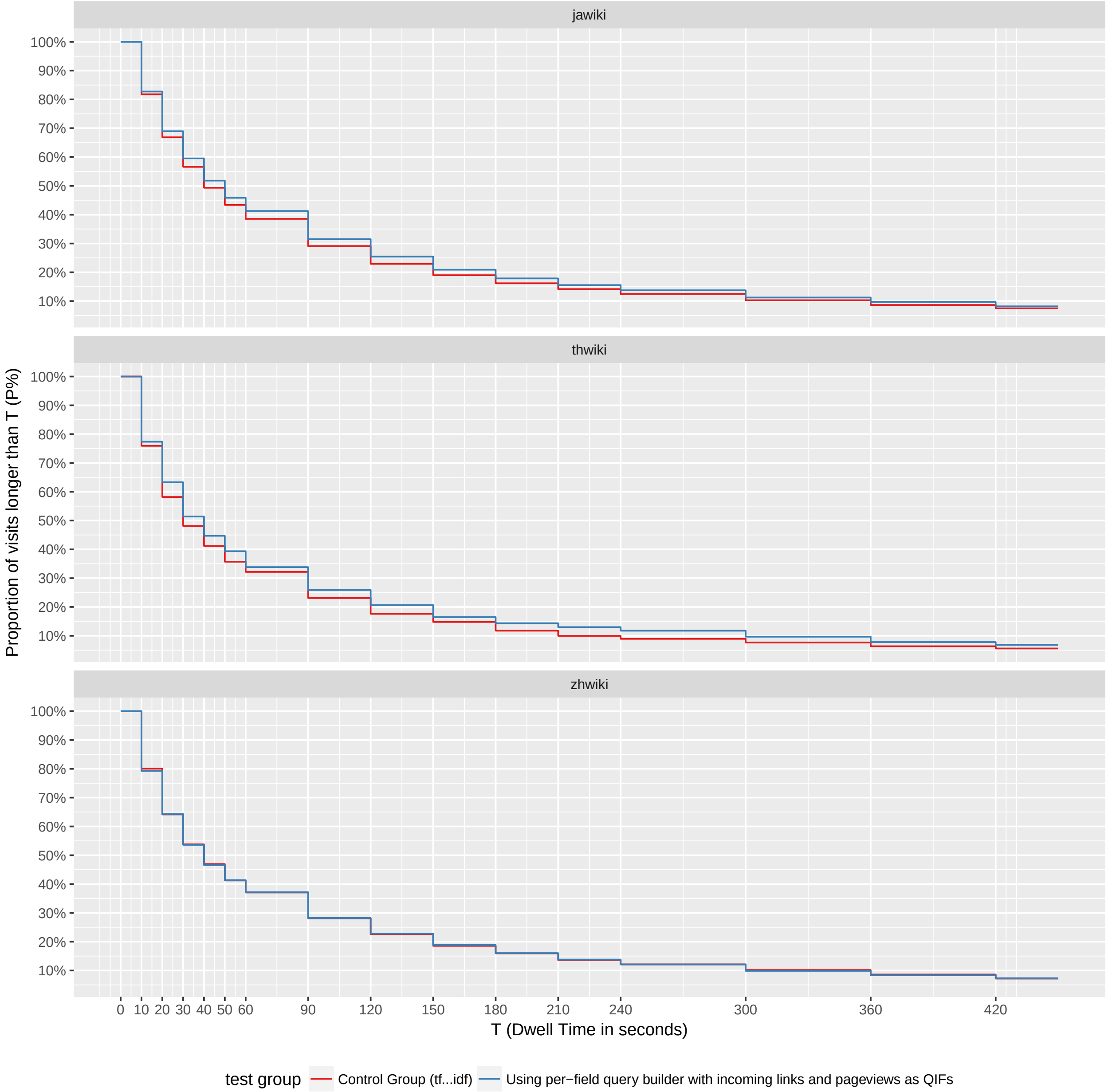

Figure 10: At time T, at most P% of users still stay on their visited pages. Broken down by test group and wiki.

Scroll

In Figures 11 and 12, we can see that users in the test group are more likely to scroll on the visited pages, but the differences are not statistically significant.

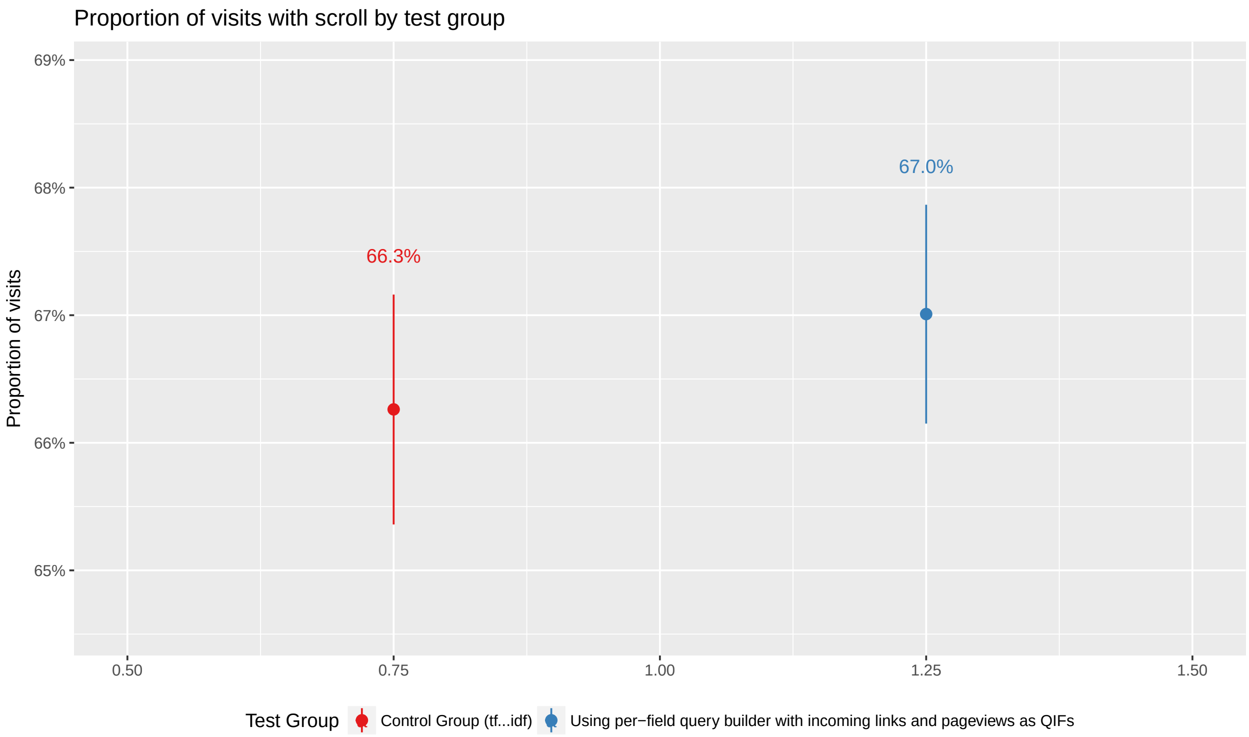

Figure 11: Proportion of visited pages with scroll by test group.

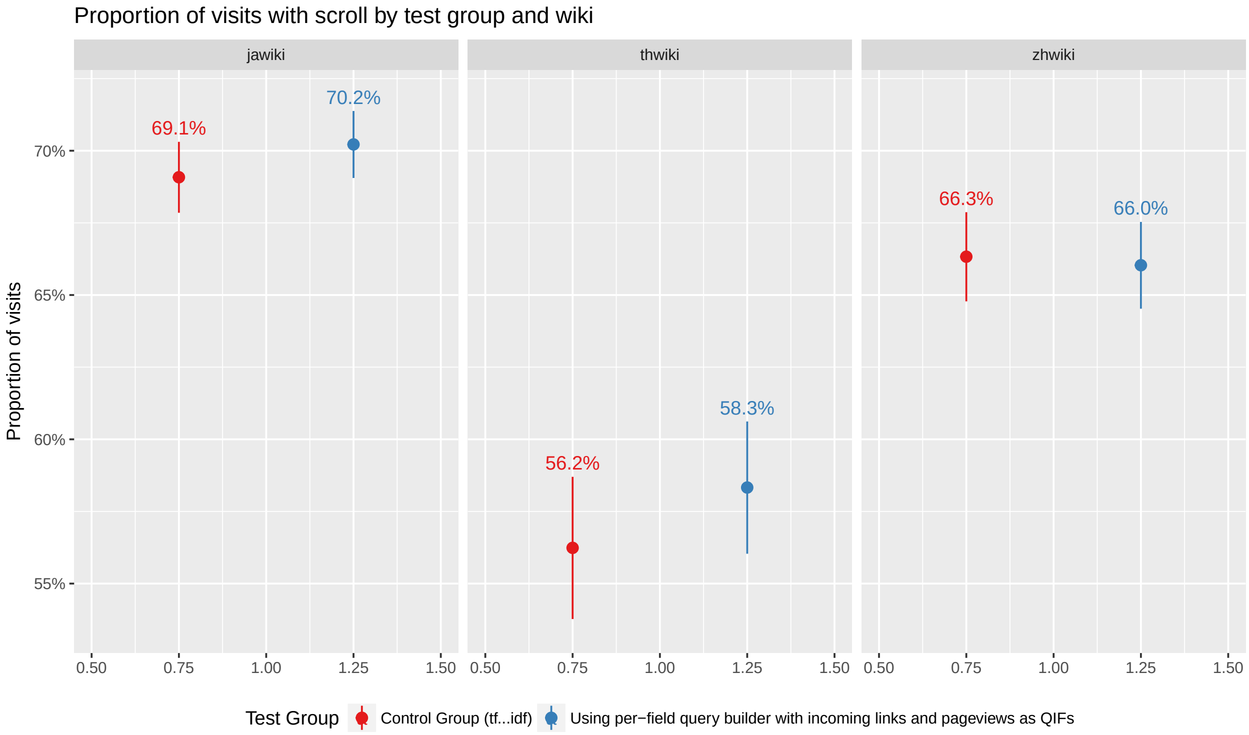

Figure 12: Proportion of visited pages with scroll by test group and wiki.

Query Reformulation

First, we tokenized queries from zhwiki, jawiki and thwiki with jieba, tinysegmenter and elasticsearch termvectors api separately. Then we filter out stop words.

We consider two queries as a reformulation if 1) they are from the same search session and share at least one result, or 2) they are from the same search session and share at least one word.

We grouped connected searches together using the rules above, then we have 51354 total search groups.

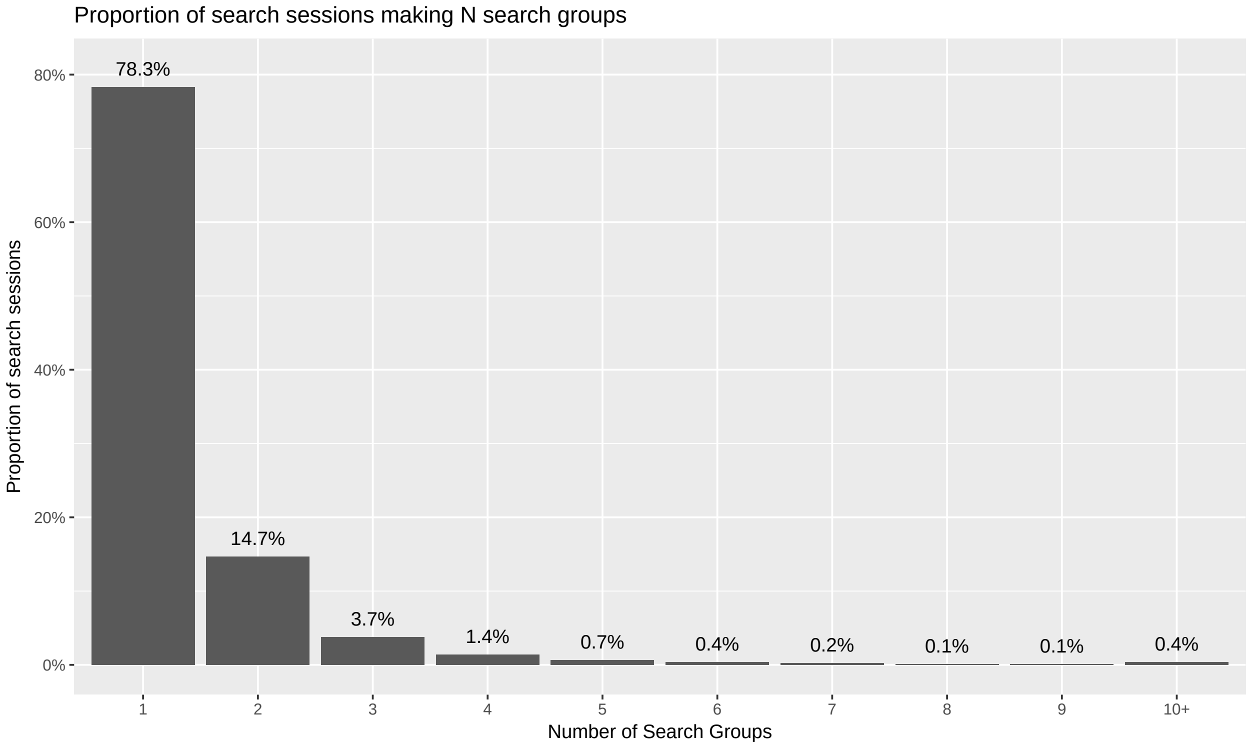

Figure 13: Proportion of search sessions making N search groups. A search group is a group of searches from the same search session, in which one search is connected with at least another one if they share at least one common word or at least one common result.

In Figure 14 and Figure 16, we can see that test group users are less likely to reformulate their queries. Figure 15 and Figure 17 show that zhwiki had the largest discrepancy between control and test group.

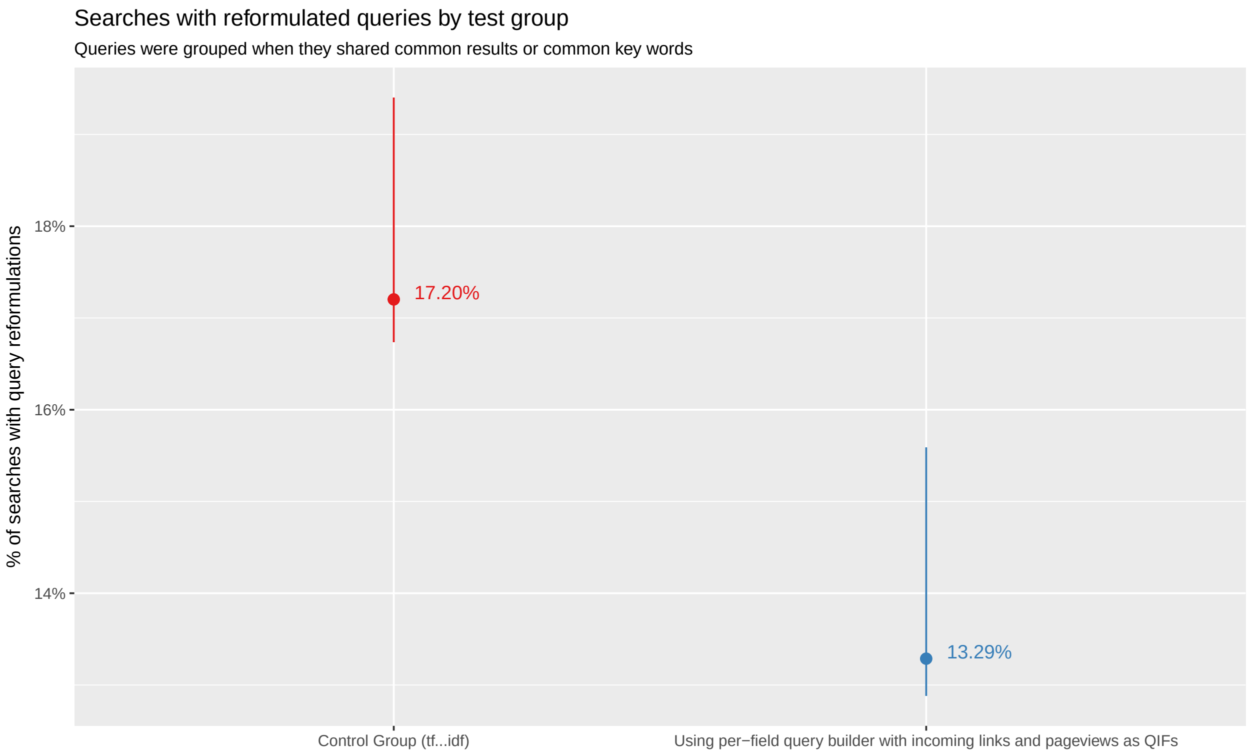

Figure 14: Proportions of searches where user reformulated their query.

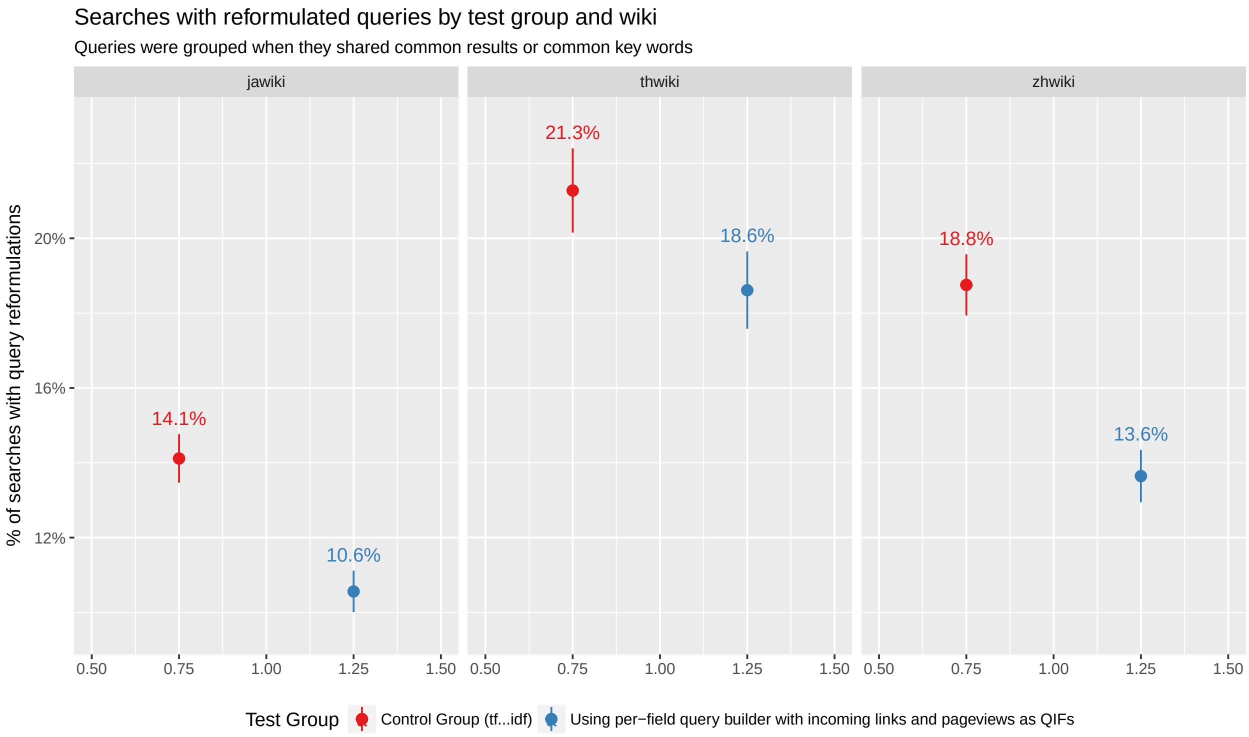

Figure 15: Proportions of searches where user reformulated their query. Broken down by wiki.

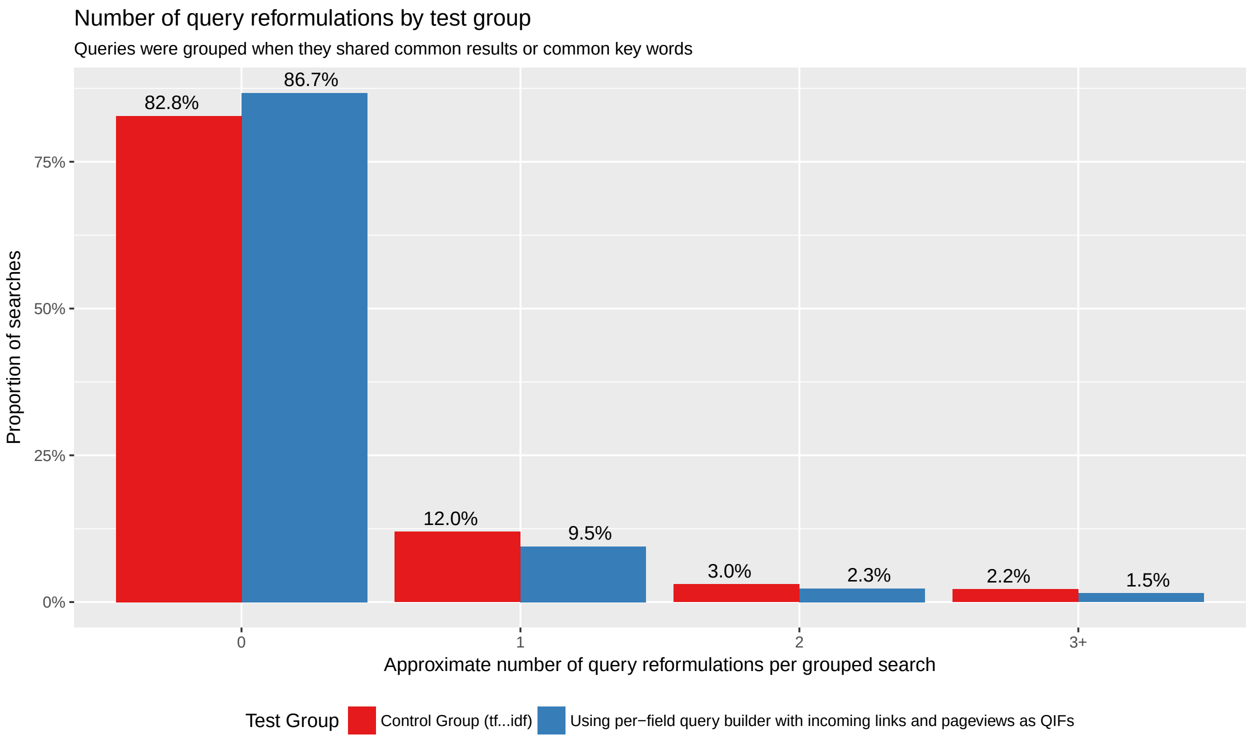

Figure 16: Proportions of searches with 0, 1, 2, and 3+ query reformulations.

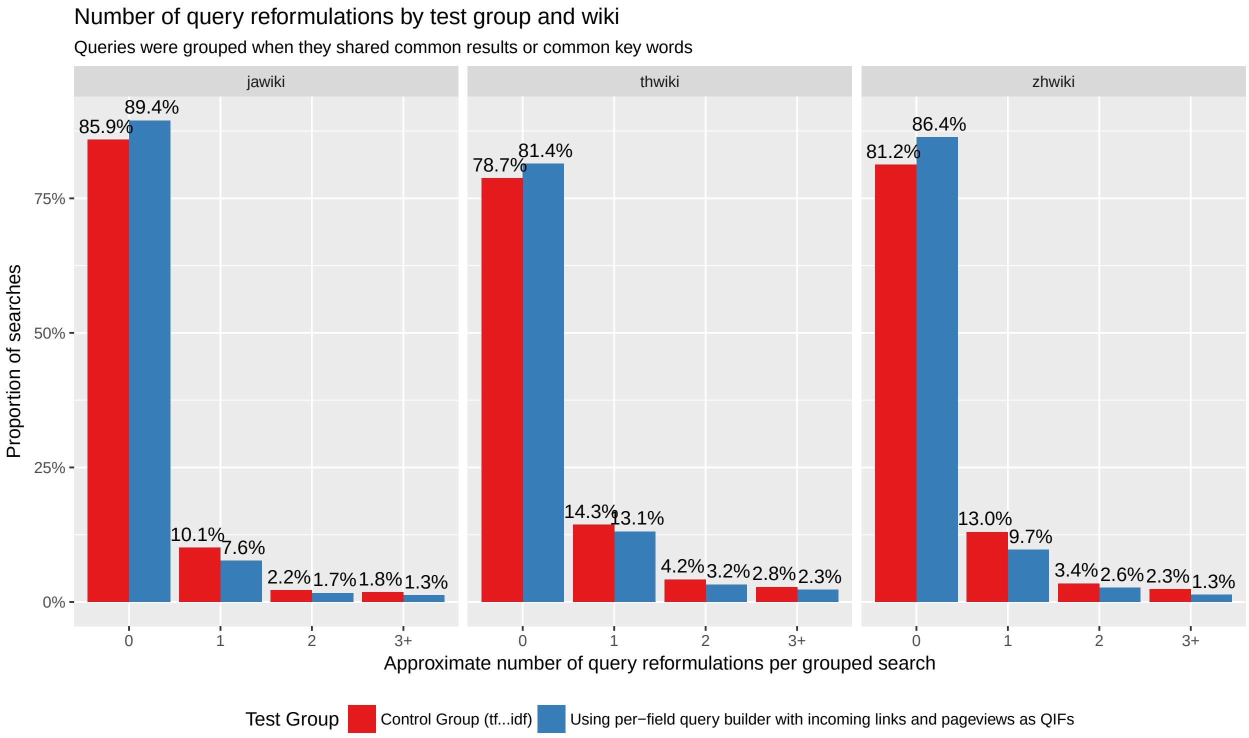

Figure 17: Proportions of searches with 0, 1, 2, and 3+ query reformulations. Broken down by wiki.

Conclusion and Discussion

For the test group, we observed a much better ZRR but slightly worse PaulScores, large decrease in clickthrough rate, and fewer users clicked on the first result first, indicating that we are showing users worse results. However, dwell-time and query reformulation analysis show that users may like the results they are getting in some aspect. We recommend deploying BM25 for all wikis but not reindexing projects in space-less languages for now.

This test revealed the performance of BM25 is unsatisfactory on space-less languages: some CirrusSearch components are dependent on the presence of spaces. We agree that the current behavior (forcing a proximity match on every query) is far from ideal but given the outcome of the test we decided not to move forward without prior work on space-less languages. We need better tokenization, and we need to track and fix all the components that make the bold assumption of the presence of spaces to activate/deactivate features.

The relatively large decrease in engagement and relevancy for zhwiki may be the result of tokenizer behavior. Chinese is the sole language in this test where we do not have a custom analysis chain. We emit only unigrams, so any page that randomly has all the same characters as the query in it will be returned. This can greatly decreases ZRR without any increase in search result quality or relevance. In English this would be roughly similar to returning any page that had all the same letters as the query.

The query reformulation analysis in this report is not ideal. Firstly, finding out shared tokens could not detect reformulated queries when users fix a typo in space-less languages. For example, when users modify their queries from “灯龙” to “灯笼” (lantern) they are fixing a typo, but tokenizers would take them as two different words. Secondly, we found that many users like to try their queries in different languages. Without enabling search across wikis in different languages, we are unable to detect this kind of reformulation.