# Query Features and Search Performance

Chelsy Xie (Analysis & Report)

Deb Tankersley (Product Management)

Mikhail Popov (Review)

Erik Bernhardson (Review)

Trey Jones (Review)

David Causse (Review)

14 November 2016

#### **Executive Summary**

Zero results rate (ZRR) – the proportion of searches that yield zero results – is a metric to measure the performance of our search system. In May 2016, we performed an analysis on zero result rate and query features using random forest and logistic regression model. This lead to us identifying question marks as the most important predictor of whether a query will yield zero results and lead to us stripping question marks from queries. With this analysis, we wanted to see which features float up to the top now after eliminating the question mark. Furthermore, we also joined search satisfaction event logging data with our Cirrus search logs to investigate the relationship between query features and other search performance metrics: clickthrough rate and PaulScore.

We used random forest and generalized linear model with elastic net penalty to shed light on the relationship between query features and search performance metrics. For ZRR, we found that whether the query has an even number of double quotes, and whether it is only punctuation and spaces are more important than other features when predicting zero results. For clickthrough rate and PaulScore, we found that query features have very small predicting power.

### Data

A user has a 1 in 200 chance of being selected for search satisfaction tracking according to our TestSearchSatisfaction2 #15700292 schema. We extracted the full-text web searching event logging data (excluding known automata) across all Wikimedia projects and languages from September 1st to October 23rd, and joined with CirrusSearchRequestSet. See data\_wTSS2.R for more details.

We used a user-defined function (UDF) for deconstructing search queries into various features in Hive. The UDF detects a variety of features such as: odd or even number of quotation marks, logical operators (e.g. AND, OR), "prefix:" or "insource:", and wildcards. For a full list of features, please see T118218 and SearchQuery.java and SearchQueryFeatureRegex.java source code.

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 (SERP) events that have the exact same search query, and then connected click events together based on the SERP connectivity. After de-duplicating, we collapsed all SERP and click events into 734,140 searches. See refine wTSS2.R for more details.

In this analysis, we used three search performance measures as our target variables: zero results rate, clickthrough rate and PaulScore. Zero results rate (ZRR) is the proportion of searches that yielded zero results (smaller is better, up to a point). Clickthrough rate is the proportion of searches that has at least one click on SERP (bigger is better). PaulScore is 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.

#### Methods

We used random forest and generalized linear model (GLM) with elastic net penalty to investigate the relationship between query features and search performance metrics. Random forest is an ensemble classification algorithm, which is known to be good at dealing with large high-dimensional datasets with many categorical features, and is less prone to overfitting. It also enables us to assess how important certain features are (through various variable importance measures) in classification. To compare with random forest, and to assess the magnitude and direction of a feature's impact, we also use logistic regression on predicting zero result and clickthrough, and linear regression on predicting PaulScore. We use the elastic net regularization when fitting the GLM, which allows for learning a sparse model where few of the weights are non-zero – like Lasso – while still maintaining the regularization properties of Ridge. We employed 10-fold cross validation to find the optimal value of the penalty parameter.

For the two classification problems – predicting zero result and clickthrough, both algorithms classify almost all data points into a single class because of the imbalanced data. To solve this problem, we performed down-sampling (deleted instances from the over-represented class) for logistic regression and stratified sampling for random forest for the training set, and use multiple metrics to measure model performance: accuracy, confusion matrix, and area under ROC curve (AUC). Although down-sampling decreases the overall accuracy of the test set, we obtained a relatively balanced class error rate and higher AUC. It is worth noting that in this analysis, instead of prediction, our goal is to figure out the feature importance for zero results and zero clicks, so that we can highlight some features for potential work and improvement. Hence, the cost of wrongly classifying some results (clickthrough) as zero result (zero click) is less than the cost of wrongly classify zero result (zero click) as some results (clickthrough), and we should pay more attention to the accuracy of zero result and zero click, instead of the overall accuracy.

For random forest, we use the Mean Decrease Gini (MDI) and Mean Decrease Accuracy (MDA) to find the features which have the most impact on classification. MDA works such that if a variable  $X_j$  is associated to response Y, then randomly permuting its values should result in a substantial decrease in the accuracy of the classification. Gini importance measures the average gain of purity by splits of a given variable. If the variable is useful, it tends to split mixed labeled nodes into pure single class nodes. Splitting by permuted variables tends neither to increase nor decrease node purities. Permuting a useful variable tends to give relatively large decrease in mean Gini-gain.

The design matrix consists of 21 dummy query features and 3 standardized continuous variables (number of terms, number of characters, and number of features detected). We considered including language (parsed from wiki ID) as a predictor but early parameter tuning tests showed that a very small increase in prediction accuracy was not enough to offset the time it would take to train. A random 80% subset of the data was used to train the model and the remaining 20% was used to test and measure the model performance.

## Results

### **Exploratory Data Analysis**

#### Proportion of searches with zero results by query feature Search queries made on the web, excluding known automata is just \* -0.0% of 3 0.0% of 5 is empty -4.1% of 242 is intitle has logic inversion (!) -7.1% of 14 is simple -11.3% of 589K ends with ? -15.6% of 1.5K has NOT -20.0% of 15 has logic inversion (-) -20.6% of 447 has non-ASCII -20.7% of 139.3K 22.2% of 36 is insource -Feature 24.8% of 2.1K has wildcard -26.9% of 26 is fuzzy search has AND -28.3% of 240 has one double quote -29.7% of 296 30.9% of 304 has odd double quotes has even double quotes -31.8% of 3.3K 35.0% of 511 is prefix -38.5% of 65 is incategory -49.3% of 71 has OR -61.0% of 118 is only punctuation and spaces has @ -78.9% of 199 50% 60% 70% 0% 10% 20% 30% 40% 80% 90% 100% Zero Results Rate Number of Queries 10 100 1K 10K 10K

Figure 1: Zero results rate by extracted features.

#### Proportion of searches with clickthrough by query feature

Search queries made on the web, excluding known automata

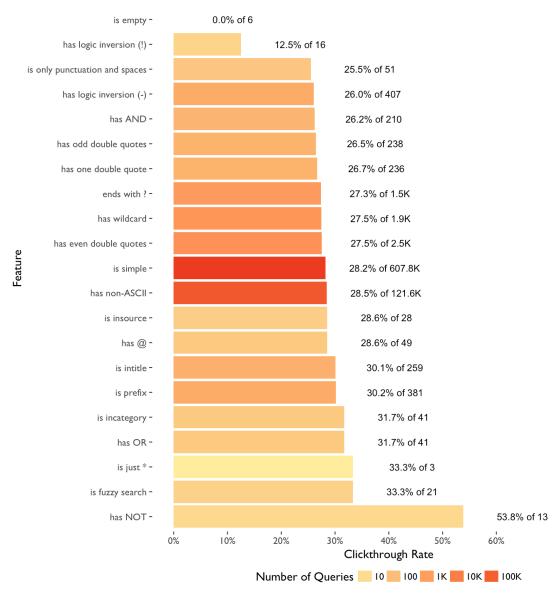


Figure 2: Clickthrough rate by extracted features.

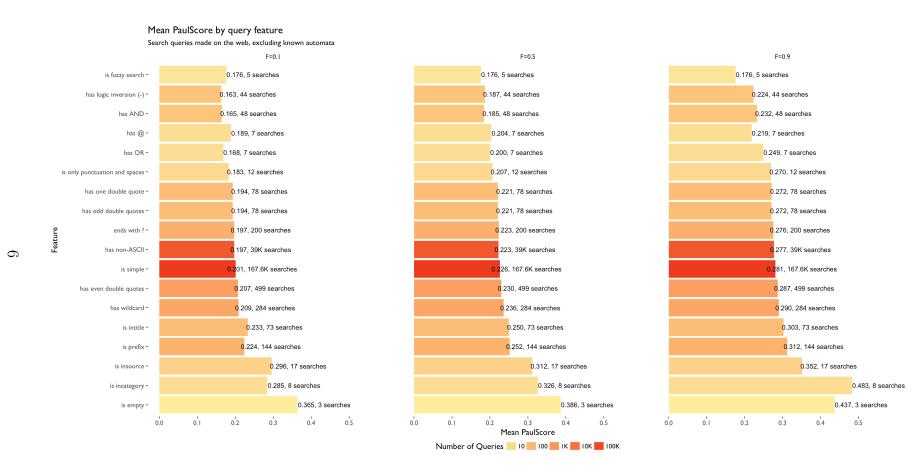


Figure 3: Mean PaulScore by extracted features. Scoring factors are 0.1, 0.5 and 0.9 respectively.

#### Zero Result Rate

**Random Forest**: Overall accuracy of test set is 0.7056147, AUC is 0.7172. The confusion matrix is:

	some results	zero results	class error
some results	91057	36474	$0.2860011 \\ 0.3497953$
zero results	6750	12547	

The following variable importance plot (Fig. 4) shows that "has even double quotes" and "is only punctuation and spaces" are more important than others when predicting zero result.

**Logistic Regression**: The two tuned parameter of elastic net penalty are very closed to 0, which means logistic regression without penalty works best here. Overall accuracy of test set is 0.6946972, AUC is 0.6726. The confusion matrix is:

	some results	zero results	class error
some results	91441	36090	0.28299
zero results	8737	10560	0.45276

Fig. 5 compares the features using the measures from both models to reveal which features the two models agree on. As before, we should pay more attention to MDA specific to queries with zero results. We can see that "has even double quotes" and "is only punctuation and spaces" are relatively more important and make zero result more likely.

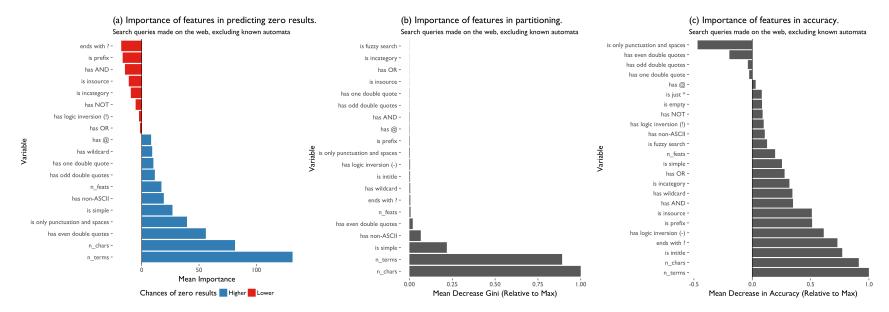


Figure 4: (a) Variable importance according to mean decrease in accuracy (increase in prediction error after permuting values of the predictor) specific to zero results queries. (b) Variable importance according to mean decrease in impurity, using Gini index. (c) Variable importance according to mean decrease in accuracy over both classes of queries (zero results and some results).

#### Variable Importance vs Logistic Regression Estimates

Search queries made on the web, excluding known automata MDA (across all queries) MDA specific to queries with some results less important more important less important more important zero results more likely zero results more likely zers results more likely and spaces zero results more likely is only punctuation and spaces 2 has a has @ is incategory is fuzzy search is incategory has even double quotes has odd bouble quotes is fuzzy search has even double quotes has logic inversion (!) has odd double quotes has non-ASCII n\_chars has non-ASCII has NOT has one double quote has logic inversion (!) has wildcard has logic inversion (-) n\_chars is insource has logic inversion (-) has one double quote n\_feats ends with? ends with? n\_terms is simple n\_feats is simple is empty is empty Logistic Regression Coefficient more important more important less important less important is intitle esults less likely zeroiseistitle less likely zero results less likely zero results less likely MDA specific to queries with zero results MDI Gini less important more important less important more important zero results more likely zero results more likely zero results more likely zero results more likely is only punctuation and spaces is only punctuation and spaces is incategory has @ is prefix is incategory has logic inversion (-) bas even double quotes is prefix has logic inversion (-) has even double quotes has non-ASCII 🕶 is fuzzy searchhas odd double quotes has AND has wildcard has odd double quotes has non-ASCII has NO fends with? n chars is insource has one double quote has logic inversion (!) n feats ends with ? n\_terms n\_terms has logic inversion (!) is simple has one double quote has OR has ÓR less important more important less important more important zero results less likely zero results less likely zero results less likely zero results less likely is intitle is intitle 0.5 -0.5 0.5 1.0 1.0 0.0 (Relative) Variable Importance via Mean Decrease Accuracy less important less important more important more important zero results less likely zero results less likely

Figure 5: A scatter map of features with respect to variable importance (via relative mean decrease metrics) and logistic regression coefficient estimates. Each of the 4 plots is divided into quadrants according to how important or unimportant the feature is in random forest classification and whether a query having the feature is more or less likely to yield zero results.

#### Clickthrough Rate

Random Forest: Overall accuracy of test set is 0.4399294, AUC is 0.5004, which means the model is not very useful. The confusion matrix is:

	zero click	clickthrough	class error
zero click	23986	61909	$0.7207521 \\ 0.2275389$
clickthrough	9444	32061	

**Logistic Regression:** Overall accuracy of test set is 0.4801099, AUC is 0.5314. The confusion matrix is:

	zero click	clickthrough	class error
zero click	35255	50640	$0.589557 \\ 0.375714$
clickthrough	15594	25911	

Because AUC is too close to 0.5, and the class error for zero click is too large, we do not think query features have large enough predictive power on clickthrough rate, and thus we are not reporting variable importance here (see appendix below).

#### **PaulScore**

**Linear Regression:** We fit linear regression models (lambdas of elastic net penalty are very close to 0) for PaulScore when scoring factor equals 0,1, 0.5 and 0.9. The explained deviance by models are very small, less than 0.3%, and R squared ranges from 0.005 to 0.009. Therefore, we do not think query features have enough predictive power on PaulScore, and thus we are not reporting variable importance here.

Random Forest: Random forest regression did a little bit better than linear regression, but the performance measures are still too small: The R squared of test set are 0.014, 0.011 and 0.0013, the mean squared errors are 0.986, 0.989 and 0.999, for PaulScore when scoring factor equals 0,1, 0.5 and 0.9 respectively. Therefore, we do not think query features have enough predictive power on PaulScore, and thus we are not reporting variable importance here.

## Appendix

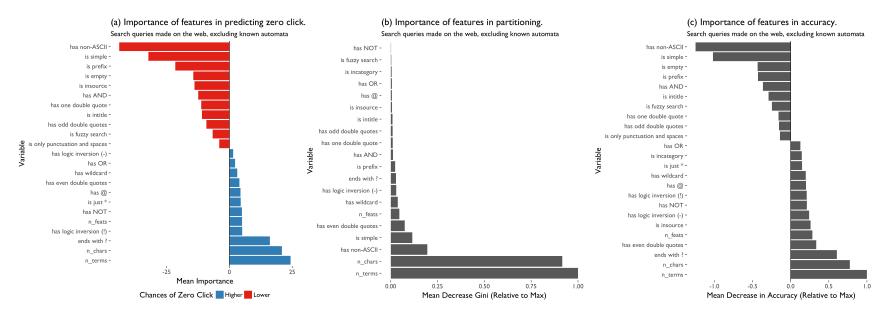


Figure 6: (a) Variable importance according to mean decrease in accuracy (increase in prediction error after permuting values of the predictor) specific to zero click queries. (b) Variable importance according to mean decrease in impurity, using Gini index. (c) Variable importance according to mean decrease in accuracy over both classes of queries (zero click and clickthrough).

#### Variable Importance vs Logistic Regression Estimates

Search queries made on the web, excluding known automata MDA (across all gueries) MDA specific to queries with clickthrough less important less important more important more important clickthrough more likely clickthrough more likely clickthrough more likely clickthrough more likely is fuzzy search is empty has AND has one double quote is insource has odd double quotes n\_chars prefix has one double quote is empty is incategory ends with ? n\_terms is fuzzy search has bdd double quotes n terms is only punctuation and spaces is only punctuation and spaces is insource -0.5 has noņ-ASCII has non-ASCII is simple has NOT has NOT is simple has logiç inversion (!) has logic inversion (!) has wildcard has wildcard Logistic Regression Coefficient has logic inversion (-) has logic inversion (-) has @ has even double quotesportant less important less important more important has even double quotes clickthrough less likely clickthrough less likely clickthrough less likely clickthrough less likely MDI Gini MDA specific to queries with zero click less important less important more important more important clickthrough more likely clickthrough more likely clickthrough more likely clickthrough more likely is incategory has OR has one double quote is intitle is insource n\_chars n\_chars n\_terms has AND n\_terms is fuzzy search is insource has odd double quotes ends with ? has AND has odd double quotes is fuzzy search is only punctuation and spaces has ÓR is intitle is only punctuation and spaces -0.5 has non-ASCII has NOT is simple is simple has NOT has logic inversion (!) has logic inversion (!) has wildcard -1.0 has wildcard has logic inversion (-) has logic inversion (-) has even double quotes noortant has even double quotesore important less important less important clickthrough less likely has @ clickthrough less likely clickthrough less likely clickthrough less likely has @ (Relative) Variable Importance via Mean Decrease Accuracy

Figure 7: A scatter map of features with respect to variable importance (via relative mean decrease metrics) and logistic regression coefficient estimates. Each of the 4 plots is divided into quadrants according to how important or unimportant the feature is in random forest classification and whether a query having the feature is more or less likely to yield at least one click.

less important
clickthrough less likely
a less important
clickthrough more likely
a more important
clickthrough less likely

more important