Performance of *count* in result sets that exceed 10,000 records

Brian Stucky

April 15, 2014

Objective: The purpose of this analysis was to characterize the behavior of the *count* value for VertNet search result sets that exceed 10,000 records. We now know that *count* should always be correct for result sets of 10,000 records or less, but whether or not *count* provides a useful representation of true result set sizes for larger searches was not known. This, in turn, will allow a decision as to how *count* should be reported to the end user for large result sets.

Methods: I assembled a set of 16 test queries with result sets that all exceeded 10,000 records. The number of matching records for these queries ranged from 12,851 to 278,484, with the remaining queries' result sizes distributed throughout this range. Queries varied in complexity from simple queries matching only on institution name to more complex queries that included taxonomic keywords or Boolean operators. Each query set was run twice on the VertNet search API for a total of 32 data points. For each query, the true number of results was counted by iteratively paging through the result set using App Engine cursors.

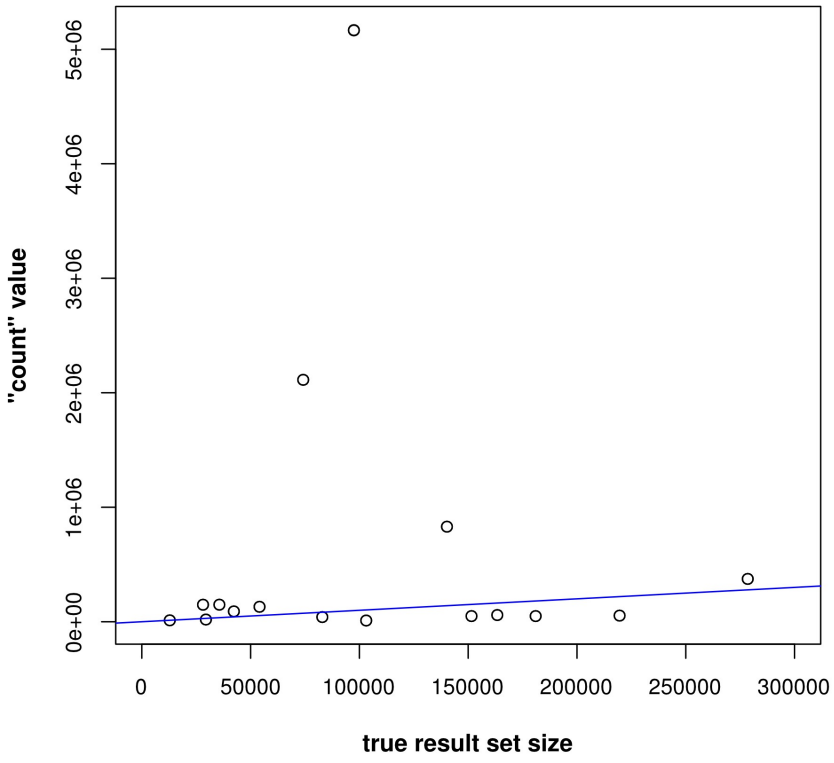

Figure 1. Scatterplot of the value of count *versus the true result set size. The solid blue line indicates the identity line (i.e., y = x). If error estimates were reliable, all points would cluster about this line.*

Results: Figure 1 depicts the relationship between *count* and the true result set size for these data. There seemed to be little discernible pattern to the data. Some *count* estimates were relatively close to the true result set size, while others were in error by an order of magnitude or more.

Furthermore, the largest result sets did not generally correspond with the largest *count* errors. For example, the single worst result was for a query that returned 97,467 records and for which App Engine estimated a result set size of 5,167,208 records. In contrast, two queries that returned more than 200,000 records consistently had more reliable *count* estimates.

To further explore the relationship between *count* and the true result set size, I fit a simple linear model to the data that attempted to predict the true result set size from the value of *count*. The resulting model had no predictive power and was not statistically significant ($R^2 = 0.000685$, $p = 0.887$). Examining the data scatterplot indicates that the model's failure is not a problem of misspecification, but instead reflects a true lack of meaningful relationship between the variables.

Conclusions: These results strongly suggest that for result sets of more than 10,000 records, *count* contains almost no consistently useful or reliable information. In numerous cases, *count* would not even be able to provide an order-of-magnitude estimate. Consequently (and unfortunately), it seems that the best course of action might be to simply return ">10000" or something similar whenever *count* exceeds 10,000.