A difference of a factor of 70,000 between hit counts and results returned in Google

Ernest Davis
Dept. of Computer Science
New York University
New York, NY 10012
davise@cs.nyu.edu

January 20, 2015

Abstract

On January 1, 2015, the author carried out a Google search for the quoted string "it is you who are mad". The results page stated at the top that there were "About 2,840,000 results", but in fact Google returned only 40 results. Attempts at replication some days later yielded discrepancies that were similar though not as large. This confirms what has often been observed, that very great caution is necessary when using Google hit counts as a measure for any purposes.

It has been known at least a decade that hit counts returned by web search engines are quite unreliable measures (Thelwall 2008) (Lapata and Keller 2005) (Veronis 2005). The hit count given for a particular query can vary substantially, not only from one search engine to another, but across two queries to the same search engine for two different clients or for the same client at two different times.

In 2005, a number of authors reported that comparing Google hit counts for related queries gave results that were either logically inconsistent with one another or entirely implausible. For instance, Lapata and Keller reported that in 2005 the query "Chirac OR Sarkozy" returned fewer results than the query "Chirac", and the query "applesauce -aosdnao" returned more results than the query "applesauce". Veronis reported that a search for the word "the" restricted to English returned a hit count that was 1% of the hit count for the search on "the" with language unrestricted. Veronis also did an experiment in which he did Google searches on 50 English words of mid-range frequencies (accumulated, alive, ancestor, bushes etc.) (a) in the entire Web (b) restricted to English. In every case, the pages count for (b) was just about exactly 56% of the page count for (b). First, this is an impossibly small percentage; and second, it is impossible that they should all be 56%. I myself did some partial replications in 2007, and got similar results (Davis, 2007).

As of 2015, Google seems to have substantially solved these problems; at least, in a few tests carried out in preparing this note, I was unable to replicate them. However, substantial inconsistencies between hit counts for the same query taken at different times persist.

As discussed in (Satoh and Yamana 2012), hit counts are often used as statistical data in such fields as natural language processing, ontology construction, and analysis of social networks. The unreliability of the hit count numbers implies that great caution is needed in using this kind of data for such purposes. Satoh and Yamana, as well as other researchers, have studied methods for obtaining more accurate numbers.

Satoh and Yamana explain the variance over time as the results of updates to the index, inconsistency between multiple indexes, and inconsistency across different search units.

However, at least in the case of search for quoted strings in Google, the discrepancies between the hit count figure and the actual number of results returned can be enormous — much larger than seems to have been previously reported in the literature. Moreover, the results page returned by Google itself can be internally inconsistent. This kind of behavior can hardly be explained in terms of the kinds of difficulties discussed by Satoh and Yamanah.

On January 1, 2015, I carried out a Google search on the quoted string "it is you who are mad", with ten results per results page (the default). At the top of the first results page, it was stated that there were "About 2,840,000 results". However, the links to additional results pages at the bottom indicated that there were only four pages of results i.e. between 31 and 40 results. The second and third page were similar. The fourth page stated at the top that there were "About 37 results"; however, it displayed ten new results. Thus, Google in fact returned a total of 40 results; a difference of a factor of 71,000.

Unfortunately, I did not save a page image of these results pages. However, I repeated the experiment, with the same string, on January 19, 2015, with similar though less extreme results. This time Google found a total of 28 results, but on the first two pages reported that there were "About 119,000 results" a difference of a factor of 4250.

The third and final page of results states that "in order to show you the most relevant results, we have omitted some entries very similar to the 30 already displayed" and invites the user to "repeat the search with the omitted results included." Clicking on this yields a sequence of results pages that announce that there are "About 2,760,000 results" but in fact contain only 50 results, a discrepancy of a factor of 55,200. It is also noteworthy that some of the results in the second collection are not in fact similar to any of the results in the first collection, e.g. the last three entries on the fifth page, "Controlling, obsessive husband," "Wing's Mu 2014", and "Worst TV adverts of the moment."

Image of all three result pages from the first set and of pages 1 and 5 of the second set can be found at `http://www.cs.nyu.edu/faculty/davise/papers/HitCountImages.pdf`

Of course it is possible that hit count is more nearly accurate, and that it is the number of results actually returned that is misleading. But, of course, this merely changes the problem, it does not eliminate it. First, the factor of 23.9 difference between 2,840,000 hit counts and 119,000 is itself rather large. Second, it is unexplained why Google only returns 28 or 40 or 50 actual results.

These results confirm the rule that considerable caution is necessary when using Google hit counts as data for any purpose.

References

E. Davis, "And Now For Something Completely Different", class notes for *Web Search Engines,* New York University. `http://cs.nyu.edu/courses/fall07/G22.2580-001/lec14.html`

M. Lapata and F. Keller, 2005. "Web-Based Models for Natural Language Processing," *ACM Translactions on Speech and Language Processing,* **2:**1. `http://homepages.inf.ed.ac.uk/mlap/Papers/tslp05.pdf`

K. Satoh and H. Yamana, 2012. "Hit Count Reliability: How Much Can We Trust Hit Counts?" *APWeb 2012* and *Lecture Notes in Computer Science 7325*, pp. 751-758. `http://dl.acm.org/citation.cfm?id=2259221`

M. Thelwall, 2008. "Quantitative Comparisons of Search Engine Results", *ASIS& T.* Wiley InterScience. `http://onlinelibrary.wiley.com/doi/10.1002/asi.20834/full`

J. Veronis, 2005. "Google's missing pages: Mystery Solved?" *Technologies du Langage,* (blog) http://blog.veronis.fr/2005/02/web-googles-missing-pages-mystery.html