

# Usage Number of Pages Found by Search Engines for Business Analytics

Alex Zolotovitski, Richard Castro  
StatVis Consulting, San Diego, CA, US  
[Alex.Zolot@statvis.com](mailto:Alex.Zolot@statvis.com)

**Abstract** - In this paper, we describe the usage of number of web pages found by search engines for business analysis. To realize this approach we created the GoogMeter application ([www.googmeter.com](http://www.googmeter.com)) that delivers and analyzes the data. GoogMeter inputs lists of any objects and properties and produces and analyzes contingency tables of total number of found pages having the combination objects and properties taking in account their proximity in text.

This method provides valuable source of statistical information related to any objects for cross-sectional and time series analysis. We compare results obtained by this method with usage of number of searches produced by Google Insights for Search.

**Keywords:** Web maps, Business Analytics, Googmeter, Google Insights.

## 1 Introduction

Many companies spend a lot of time and money to obtain data related to their customer behavior. Usually, to get knowledge from the Internet we use a two stage process:

1. Google to get a list of links;
2. Follow the links to specific web sites for requisite information.

At the same time huge amount of data can be extracted directly from Internet using search engine statistics. There are two ways to obtain knowledge about any objects directly, in one stage, because there are two important numerical characteristics for any web search query:

- 1) The number of searches with the query.
- 2) The number of page found by a search engine and

The first approach is used in the Google Insights for Search ([www.google.com/insights/search](http://www.google.com/insights/search)) that shows the total number of searches done on Google over time. Using it,

Michael Cavaretta showed recently [2], that number of customers' searches correlates with their purchase behavior.

To realize the second approach invented by Douwe Osinga[1], we created GoogMeter application ([www.googmeter.com](http://www.googmeter.com)) that analyzes number of pages found by search engines for combinations of words taking in account their proximity in text.

Figuratively speaking, Google Insights measures the number of questions asked by customers and GoogMeter measures the number of answers in the Internet.





### GoogMeter

GoogMeter ([www.googmeter.com](http://www.googmeter.com)) is a web comparator that measures Internet proximity between any objects (Obj) and properties (Prop). The most difficult and important tasks are to ask good questions and to analyze and interpret the answers.





As the simplest example, let we do not know which parties John McCain and Barack Obama belong. Then Googmeter gives as:

Objects: McCain, Obama  
Properties: democratic, republican  
Search Engine: Google

### Number of Pages Found, ths

	democratic	republican
McCain	26800.0 	28700.0 
Obama	27900.0 	28700.0 

Indexes (ratios of actual numbers of found pages to expected numbers):

	democratic	republican
McCain	99.0 	101.0 
Obama	101.0 	99.0 

 ~ Number of pages  = Yes  = No

Figure 1. The simplest example of Googmeter’s output.

and now we see, that McCain is associated with the republican party while Obama to the democratic party. In spite of the fact, that numbers of pages found with both properties are very close, usage of indexes gives us a correct result.

GoogMeter uses the following algorithm: it runs the search queries with all combinations of Objects and Properties on the specified search engine, gets the number of pages found  $N(\text{Obj}, \text{Prop})$  and creates contingency tables where rows correspond to Objects and columns to Properties.

From the tables it calculates Totals by columns -  $\text{Tot}(\text{Obj})$ , rows -  $\text{Tot}(\text{Prop})$  and overall  $\text{Tot}$  and then empirical probabilities

$$p(\text{Obj}) = \text{Tot}(\text{Obj}) / \text{Tot}, \quad p(\text{Prop}) = \text{Tot}(\text{Prop}) / \text{Tot}.$$

After it we obtain the expected number of pages

$$E(\text{Obj}, \text{Prop}) = \text{Tot} * p(\text{Obj}) * P(\text{Prop})$$

and indexes

$$\text{Ind}(\text{Obj}, \text{Prop}) = 100 * N(\text{Obj}, \text{Prop}) / E(\text{Obj}, \text{Prop}).$$

GoogMeter prints Number of found pages  $N(\text{Obj}, \text{Prop})$  and Indexes  $\text{Ind}(\text{Obj}, \text{Prop})$  and visualizes the table plotting horizontal bars or bubbles that colored green if Actual Numbers are greater than Expected and red in opposite case.

There are two variants for the bar’s width or bubble’s volume:

1. Width is proportional contribution to Chi2 statistics  $\sim (N - E)^2 / E$
2. Width is proportional  $\sim |\ln(\text{Ind}/100)| = |\ln(N/E)|$

We use bubbles rather than bars because radius of bubbles is proportional to cubic root of volume, so in case when range of values to present is very wide chart becomes more compact than in case with bars.

## 1.1 Program

In GoogMeter we used python code that is based on modified Douwe Osinga’s code[1]. Now performance of the program can be essentially improved using a new Google doc API function ImportXML().

## 2 Usage of Googmeter for Business Analytics

This data provided by GoogMeter can be used in many areas of business analytics.

### 2.1 Quality of Products

If we are interested in analysis of quality of Sun Microsystems’ products, we could use such objects as “Microsoft, IBM, Hewlett-Packard, Dell, Sun Microsystems”, and such properties as “excellent, bad, reliable, unreliable, problem, bug, failure, crash, new, troubleshooting, friendly”. It gives us the following result table for indexes:

**Indexes (ratios of actual nubers of found pages to expected numbers)**













































	IBM	Hewlett-Packard	Dell	Sun Microsystems
excellent	94.9 	95.9 	108.7 	79.0 
bad	97.8 	91.8 	105.0 	91.7 
reliable	111.2 	117.5 	83.5 	115.8 
unreliable	116.7 	107.8 	80.7 	114.9 
problem	105.9 	100.9 	93.6 	105.0 
bug	104.5 	81.0 	95.8 	128.7 
failure	109.1 	126.7 	83.7 	113.9 
crash	103.9 	99.3 	96.3 	102.3 
new	96.8 	99.2 	103.3 	98.9 
troubleshooting	104.1 	100.6 	93.0 	120.5 
friendly	93.2 	99.3 	107.8 	90.0 

Figure 2. Comparison of HW brands.

Analyzing characteristics of software we could use Objects: “vista, solaris, linux, ubuntu, red hat”

and the same Properties: “excellent, bad, reliable, unreliable, problem, bug, failure, crash, new, troubleshooting, friendly”.

It gives us the following result:

**Indexes (ratios of actual numbers of found pages to expected numbers)**

	vista	solaris	ubuntu	red hat
excellent	103.5	94.3	85.3	85.4
bad	102.7	85.1	102.2	83.0
reliable	94.5	138.9	104.9	113.9
unreliable	90.5	150.5	122.7	123.9
problem	92.6	132.9	127.1	113.8
bug	77.5	183.7	199.3	136.5
failure	88.3	165.5	119.8	135.7
crash	94.8	134.5	115.6	101.5
new	103.6	79.2	84.1	101.8
troubleshooting	90.5	144.6	115.1	139.2
friendly	108.7	68.0	74.0	69.7

Figure 3. Comparison of SW brands.

In the last example we swapped Objects and Properties to get the narrower table.

We can analyze the resulting tables using principal component analysis, SVD or to obtain the distance matrix for the objects and plot it in appropriate projection to 2D plane, as it was done by Douwe Osinga in [1], where objects were countries.

Using four graphic parameters: two axes, size and color of bubbles, we can visualize four properties of objects, e.g.:

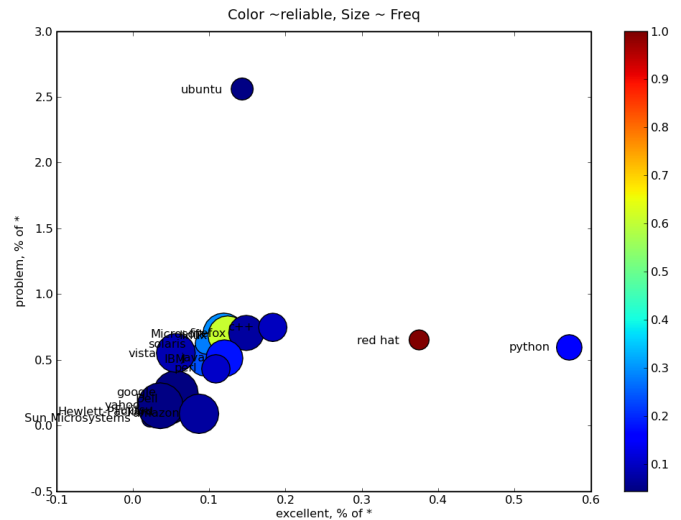


Figure 4. Comparison of SW brands with visualization of four parameters.

Figure 4 shows visualization of four properties for SW brands: “excellent, % of \*” – axis X, “problem, % of \*” – axis Y, “reliable” – color and “frequency” – size. E.g. “problem, % of \*” for “ubuntu” means ratio of numbers of pages found for queries “ubuntu problem” and “ubuntu”.

## 2.2 Marketing

The same approach can be used for marketing analysis if we choose appropriate terms as properties. Let see the degree of web association between companies on one side and target markets – countries and universities on other side.

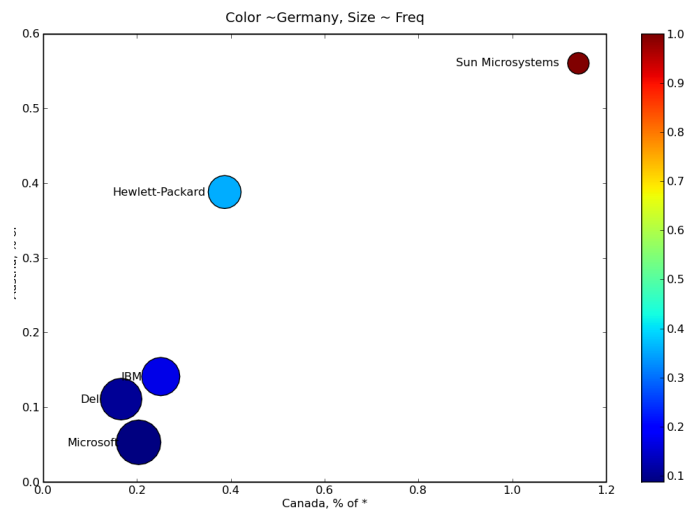


Figure 5. Comparison of HW brands by country.

Microsystems” with Canada about twice stronger, than with Austria, but association of brand HP with Canada and Austria is about the same.

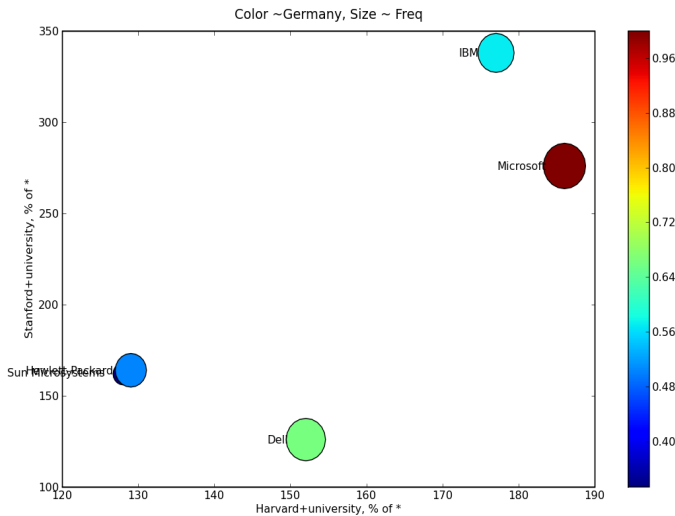


Figure 6. Comparison of HW brands by university associations.

At this charts labels on axes are “Stanford University, % of \*” and “Harvard University, % of \*”. From this chart we see that brands IBM and Microsoft are stronger related to universities than other HW brands. At the same time brands Sun Microsystems and HP are stronger related to Stanford university, but Dell – to Harvard university.

### 2.3 Time Series.

Repeating the same Googmeter queries daily, we obtained the following charts:

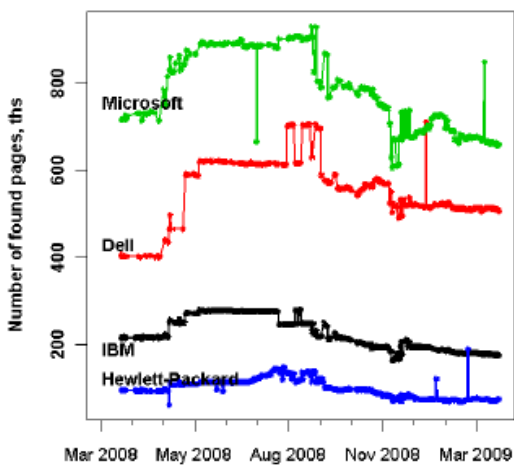


Figure 7. Time series for HW brands from Googmeter

Obviously, the number of found pages depends on internal state of search engine that changes in time, sometimes producing outliers on the chart, so only time series for

relative characteristics for different objects make sense. To get rid of this effect we will smooth the time series lines excluding jumps more than 10% - the resulting time series are plotted on Figure 8.

It is interesting to compare the above chart with chart produced for the same objects by Google Insights:

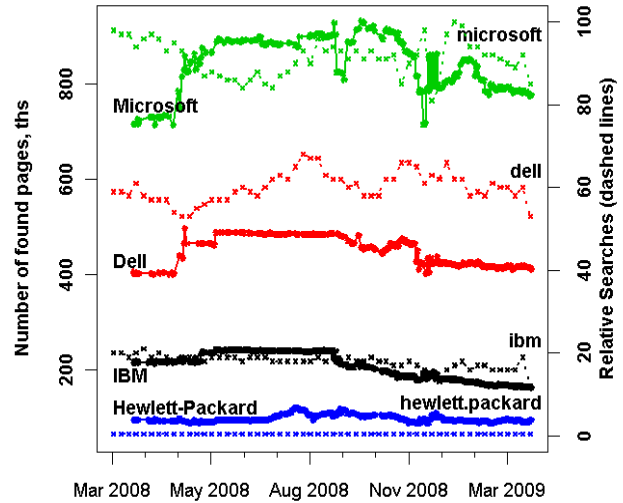


Figure 8. Time series for HW brands. Comparison between smoothed Googmeter (left axis, solid lines) and Google Insight (right axis, dashed lines) data.

We do not see any clear correlation between the time series of number of found pages and number of searches; coefficients of correlation for Microsoft, Dell and IBM are equal  $-0.77$ ,  $0.61$  and  $-0.24$ . Probably our period of observations (12 months) was too short to reveal this correlation, because on long time scale (years) such positive correlation must have place, but we can expect a time lag between number of searches and number of pages found.

At the same time discrepancy between directions of the trends could indicate necessity of business actions. Ratio of number of searches to number of pages found could be related to number of visitors of the pages related to a term.

Internet behavior of customers closely related to other their commercial activity, e.g. Figure 9 shows obvious correspondence between number of searches and volume of stock trades.

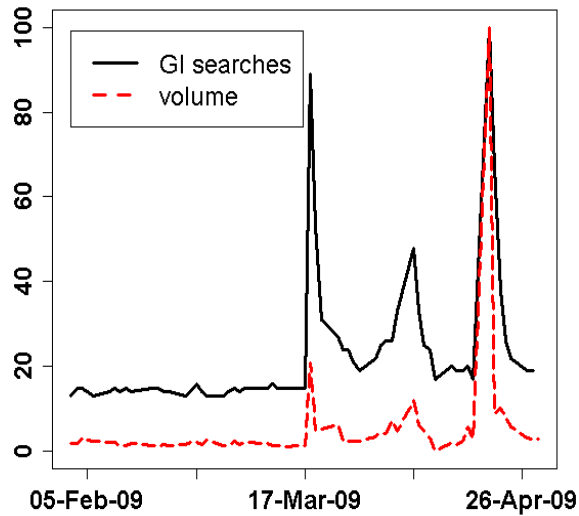


Figure 9. Time series for Google Insights number of searches for “Sun Microsystems” (solid line) and volume of stocks trade with “JAVA”(dashed line).

### 3 Conclusions

We see that analysis of data about number of found pages and number of searches that could be obtained from Googmeter and Google Insights can give valuable business information.

### 4 References

- [1] [Douwe Osinga](http://douweosinga.com/projects/mappedweb) Mapped Web  
<http://douweosinga.com/projects/mappedweb>
- [2] Michael Cavaretta, Sales Forecasting Using Google Searches. SAS 2008 Data Mining Conference, Las Vegas,  
[www.sas.com/events/dmconf/abstract.html#cavaretta](http://www.sas.com/events/dmconf/abstract.html#cavaretta)