Googlewhack and Internet Search Result Probabilities, Jonathan Lansey, Advisor: Bruce Bukiet, NJIT

Introduction

We study the number of internet search results returned from multi-word queries based on the number of results when each of the words are searched for individually, A, B and C.

Uniform Model

A Googlewhack is a search for two words on Google that returns exactly 1 result, as shown in **Figure 1**. **Figure 2** plots A and B vs. A/B for 351 pairs taken from Googlewhack.com. If we assume that the probability of finding a word on a given page is the same for all pages, with the total number of those pages in Google's index defined as I (as returned by a wild card search **), then the dashed lines in **Figure 2** represent for each value of A/B the values of A and B that lead to the maximum probability for a Googlewhack, given by the following equation:

$$\left(\frac{A}{I}\right)\left(\frac{B}{I}\right)I = \frac{AB}{I} = 1 = \text{Results} \equiv \text{R}$$

The exact probabilities can be derived from combinatorics. This formula is used for the contour plots in **Figures 4 and 5**.

$$p(R) = \binom{I}{R}\binom{I-R}{A-R}\binom{I-A}{B-R} \Big/ \binom{I}{A}\binom{I}{B}$$

$$p(R) = \frac{A!B!(I-B)!(I-A)!}{I!R!(A-R)!(B-R)!(I-A-B+R)!}$$

Non Uniform Model

In reality, words are more likely to be found on larger pages. We model the distribution of page sizes in the web with a Zipf power law [1,2,3] $(\#\text{Words on Page } i) = K/i^{\alpha} = n$

with i ranging from 1 to I. The probability that a word will be found on a page is proportional to the number of unique words on the page. The number of unique words is calculated from Heaps law [4,5] $(\text{Unique words in text size } n) = Kn^{\beta}$

Putting the two together, we have the probability of a word appearing on page i when there is only one result for that word.

$$p(i) = \frac{k}{i^{\alpha\beta}} = \left(\frac{1-\alpha\beta}{I^{1-\alpha\beta}-1}\right)\frac{1}{i^{\alpha\beta}}$$

We make the approximation that $p(A,i) \approx Ap(i)$

which is verified by the computational model shown in **Figure 3**. We approximate a few sums with integrals and derive an effective value for I ($I_{eff\,n}$) for n=2,3 for word pair and triplet searches, respectively

$$AB\frac{(\alpha\beta-1)^2\left(I^{2\alpha\beta}-I\right)}{(2\alpha\beta-1)\left(I^{\alpha\beta}-I\right)^2} = \frac{AB}{I_{eff2}} = R,\quad ABC\frac{(\alpha\beta-1)^3\left(I^{3\alpha\beta}-I\right)}{(3\alpha\beta-1)\left(I^{\alpha\beta}-I\right)^3} = \frac{ABC}{I_{eff3}^2} = R$$

Figure 1) A Googlwhack, Biodiversified Snacking

More Googlewhack Examples:

- Fabulated Marshmellows
- Protozoic Spliff
- Slipperiest Airscrew
- Quintupling Zugzwang
- Netherworldly Mugwumps

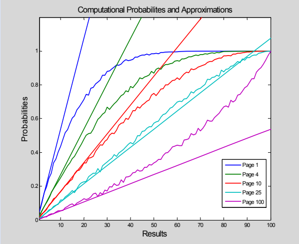

Figure 3) The probability of a word being on a given page is plotted vs. the number of total results for that word. The lines are linear approximations, good when the word appears only on a few pages. $p(A,i) \approx Ap(i)$ Data was generated from a computational model.

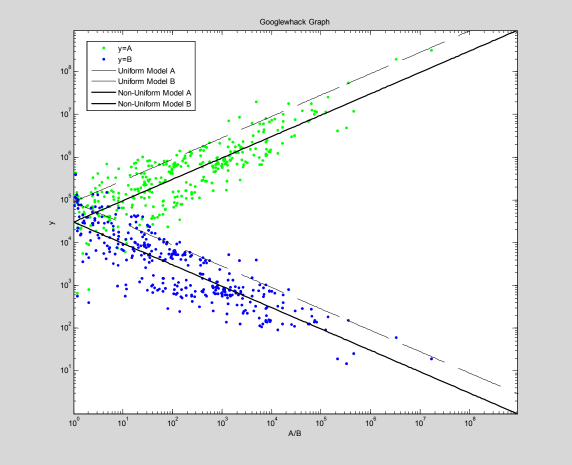

Figure 2) Plots of A and B vs. A/B for 351 Googlewhack pairs. The maximum probabilities for Googlewhacks for the uniform model and non-uniform model approximation are also plotted as dashed and solid lines, respectively. (Compare to **Figure A** below.)

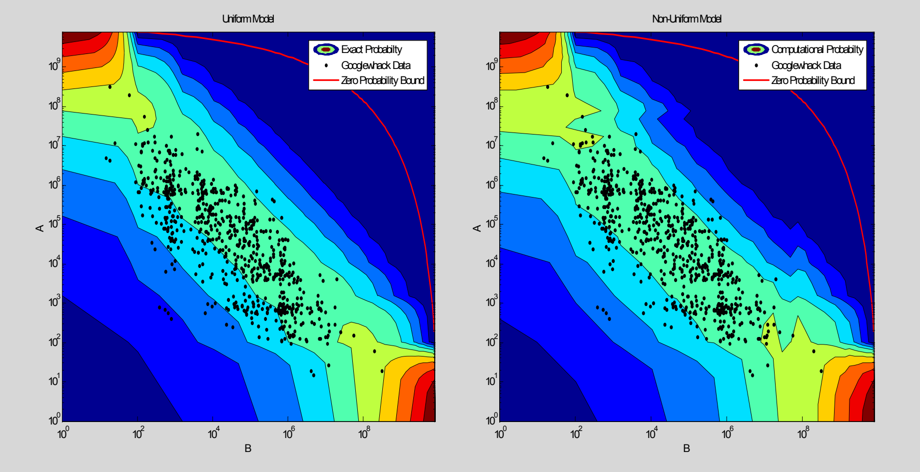

Figure 4) Left: A vs. B is plotted for each of the Googlewhacks. The contour plot is the exact probability for a given (A,B) to be a Whack. We observe that most are located near a band of high probability with a predicted shift similar to that of **Figure 2**. **Right:** The computational model is built with the assumptions discussed in "Non-Uniform Model." It is valid for all results, unlike the linear approximation. Due to obvious computing limitations, the computation was run for an index size of 30 and scaled up to 8 billion. For the sake of comparison the exact plot was also made for the same smaller index size. The red lines define a bound beyond which the probability of a Whack is zero, A+B>I+1.

We see from these log log plots that lines of equal A*B have about the same probabilities (the effect is much clearer for larger I). This motivates our choice of x axis in **Figure 5**.

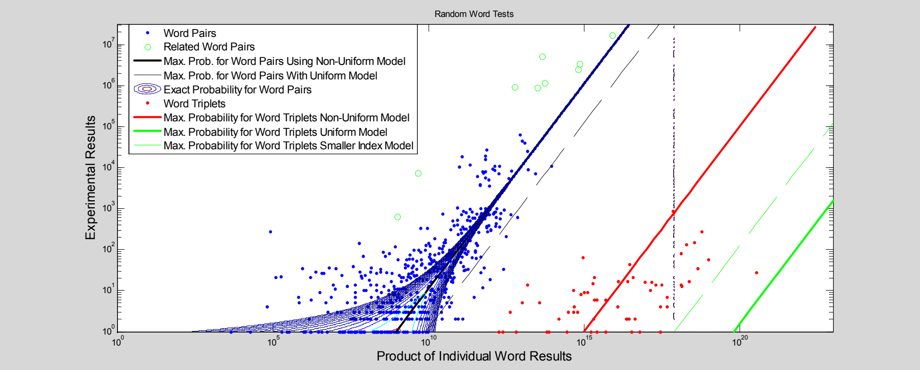

Figure 5) Each point represents a Google search: the x axis is the product of the results for each word when searched for individually and the y axis is the results when they are searched for together.

- Blue Points: Random word pairs from the Googlewhack vocabulary.
- Small Green Circles: Related words, for example: {Stairway Heaven} appears above the line with more results than expected.
- Solid Black Line: The number of results with maximum probability given A*B using the Googlewhack fitted I_{eff2}
- Dashed Black Line: The maximum probability using the true value of I
- Contours: Plots the exact probabilities with the uniform equation & Googlewhack fitted I_{eff2}
- Red Dots: Random word triplets from the Googlewhack vocabulary.
- Red Line: The max. probability for the word triplets using the non-uniform model with Googlewhack fitted I_{eff3}. The exact probabilities are non trivial to solve for three word searches.
- Solid Green Line: The maximum probability for three words using the real value of I
- Dashed Green Line: The maximum probability for three words using the value of I_{eff2}

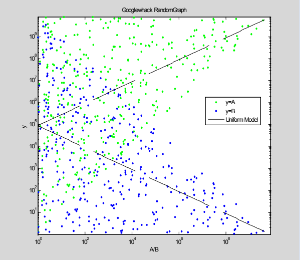

Figure A) The plot in figure 2 is slightly forced but the significance is still clear when 351 numbers (with uniformly distributed random exponents from 0 to 10) are plotted in the same way.

Conclusion

- Picking I_{eff2} to best fit the Googlewhack data in **Figure 2** gives: $\alpha\beta$=0.52. This result agrees with the accepted values for these parameters, $\alpha\approx1$ [1,2,3] and $0.4\geq\beta\leq0.6$ [4,5].
- I_{eff3} for three word searches is computed with the experimental value of $\alpha\beta$=0.52 and fits the data graphed in **Figure 5**.
- The Zipf law parameter, α, is typically easy to calculate, while measuring the Heaps law parameter, β, is very computationally intensive. If α can be measured independently, our method can easily measure β over 25 billion pages.

References

1. 1AB Downey. The structural cause of file size distributions. *SIGMETRICS/Performance*, 2001
2. LA Adamic, BA Huberman. Zipf's law and the internet. *Glottometrics*, 3, 2002,143-150
3. M Faloutsos, P Faloutsos, C Faloutsos. On power-law relationships of the Internet topology. *ACM SIGCOMM Computer Communication Review*, 1999.
4. Baeza-Yates, Ricardo and Ribeiro-Neto, Berthier. Web Usage Mining in Search Engines. *Modern Information Retrieval Addison-Wesley*, New York 1999.
5. Heaps, H. S. *Information Retrieval: Computational and Theoretical Aspects.* Academic Press, New York, 1978.

Acknowledgments

Thanks to Etan Bukiet for coding "Googlewhacker.exe," the program used for this study.

Contact

Jonathan, JCL@njit.edu
http://web.njit.edu/~jcl7/publications/googlewhack.html