Quantifying the Relationship between Hit Count Estimates and Wikipedia Article Traffic

Tina Tian, Ankur Agrawal
Department of Computer Science
Manhattan College
New York, USA

***Abstract*—This paper analyzes the relationship between search engine hit counts and Wikipedia article views by evaluating the cross correlation between them. We observe the hit count estimates of three popular search engines over a month and compare them with the Wikipedia page views. The strongest cross correlations are recorded with their delays in days. We present the results in both graphs and quantitative data among different search engines. We also investigate the predicting trends between the hit counts and Wikipedia article traffic.**

Keywords—hit count estimations; search engines; Wikipedia article traffic; cross correlation; positive delay, negative delay; prediction of Web hosting trend

I. Introduction and Related Work

When a user searches for a term, the search engines return the estimated number of Web pages related to the keyword searched, named the search engine hit count [1]. Search engine results are now widely used for measurement purposes in Webometrics. Researchers have used hit counts as input for many studies of Web information, e.g., to determine how many pages in one country link to another [2]. Cilibrasi and Vitanyi used search engine hit counts to measure semantic similarities of words [3].

Wikipedia is the largest encyclopedia in existence and it is among the fastest growing sites on the web [4]. Wikipedia has appeared in many research papers as a valuable data source of study. For example, Ponzetto and Strube used Wikipedia for computing semantic relatedness [5]. Cucerzan presents a system for the recognition and semantic disambiguation of entities based on information extracted from Wikipedia [6]. In the field of health information, Laurent and Vickers investigated whether Wikipedia article traffic correlated with epidemiological factors and compared page views statistics to a major online health encyclopedia [7].

This paper studies the correlation between search engines' hit counts and the Wikipedia article traffic. It also analyzes the predicting trends between the two resources, which can be a useful tool for business and Web publishers to promote their websites.

The rest of the paper is organized as follows. Section II describes the process of mining hit counts and Wikipedia article views and describes the approach to evaluating their correlation. In Section III, we present the results by graphing the correlations and we investigate if traffic to Wikipedia articles can predict the Web hosting trend, or vice versa. Section IV concludes the paper and proposes future work.

II. Methods

The experiment was based on 400 popular search terms from four different categories, including medicine, people, science and technology. We collected 100 terms for each category. Each term was queried against the official site of Wikipedia article traffic statistics [8], which measures the page views of a given article in a given month. Besides the total number of views in a month, the site also provides daily views in JSON format [8]. A program was developed to extract the Wikipedia page views over a period of one month and to store them in a database.

In the meantime, each search term was sent to popular search engines to collect their daily hit count. Three search engines were selected in this research, including Google, Yahoo! and Bing. Programs were built to retrieve search engines' hit counts through their APIs. Google's JSON/Atom Custom Search API returns search results of a term through RESTful requests from Google Custom Search [9]. The returned results are presented in JSON or Atom format, which can be parsed with a program.

Bing's hit counts were retrieved in a similar way using the Bing Search API, which returns the search results in XML or JSON format [10]. We used Bing's results to represent both Bing and Yahoo!, since the latter is powered by Bing's search engine [11].

We queried the search terms against the search engines' APIs and mined their Wikipedia page views. The results were observed over a month and were stored in a database. Outliers with 0 hits from a search engine were removed.

In order to analyze the relationship between the search hit counts and the Wikipedia article traffic, the cross correlation of the two was calculated. Cross correlation is a standard method of estimating the degree to which two series are correlated [12]. It is a measure of similarity of two series as a function of the lag of one relative to the other. Given two N-element series $x[i]$ and $y[i]$ where i=0,1,2...N-1, the cross correlation r at delay d is defined as in equation (1),

$$r(d) = \frac{\sum_i((x[i]-mx)(y[i-d]-my))}{\sqrt{\sum_i(x[i]-mx)^2}\sqrt{\sum_i(y[i-d]-my)^2} \qquad (1)$$

where *mx* is the mean of series *x* and *my* stands for the average of series *y*. We used the *x* series to represent the search engine's monthly hit counts of a search term and we used the *y* series for the number of views of the according Wikipedia article. Cross correlations have been observed with different days of delays assigned, ranging from -*N* to *N*. The maximum cross correlation was then selected and stored together with its delay for further analysis.

The pseudocode below illustrates the procedure of calculating the strongest cross correlation, knowing the search engine's monthly hit counts of a term and its Wikipedia monthly traffic.

```
MAX_CROSS_CORRELATION(x[], y[]){
  max_r = 0;
  FOR (delay = -N; delay <= N; delay++){
    Calculate the cross correlation r at delay
    IF (r > max_r){
       max_r = r;
       max_delay = delay;
    }
  }
  return [max_r, max_delay];
}
```

A cross correlation with 0 delay has the same value as the Pearson correlation coefficient. A negative delay represents a delay in the search engine hit count, while a positive delay stands for a delay in the Wikipedia article traffic. For example, Figure 1 shows Google's monthly hit counts for search term "Jared Cohen." Figure 2 plots the series of the monthly views of article "Jared Cohen."

The cross correlation series with a maximum delay of 30 is shown in Figure 3. One can observe that the strongest correlation occurs at delay of about 1. In other words, the series of Google hit counts from day 1 correlated with the Wikipedia article traffic from day 2. Thus, for search term "Jared Cohen," the decline of Google's hit count predicted the traffic of the same titled article on Wikipedia.

Java was used to extract hit counts from search engines and page views from Wikipedia. We wrote programs in C++ to calculate and store the cross correlations between the two. At the time of writing, Google's Custom Search API provides 100 search queries per day for free with additional queries at a cost [9]. The Bing Search API charges if there are more than 5,000 transactions per month [10].

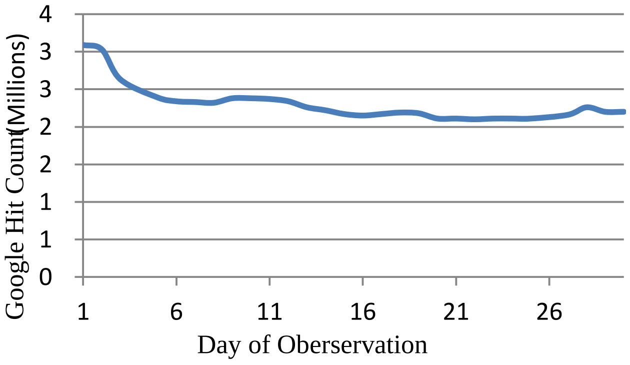

Fig. 1. Monthly Google hit counts for term "Jared Cohen"

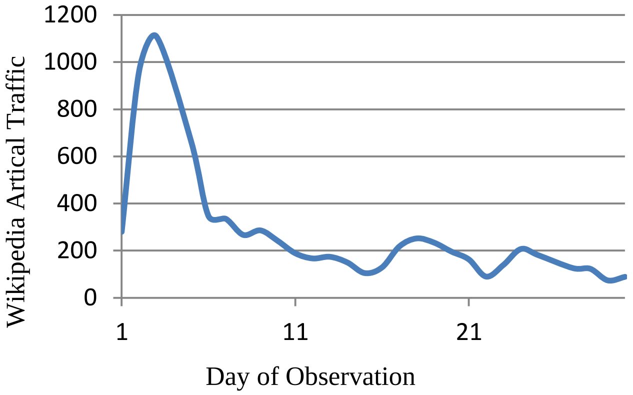

Fig. 2. Monthly Wikipedia page views of "Jared Cohen"

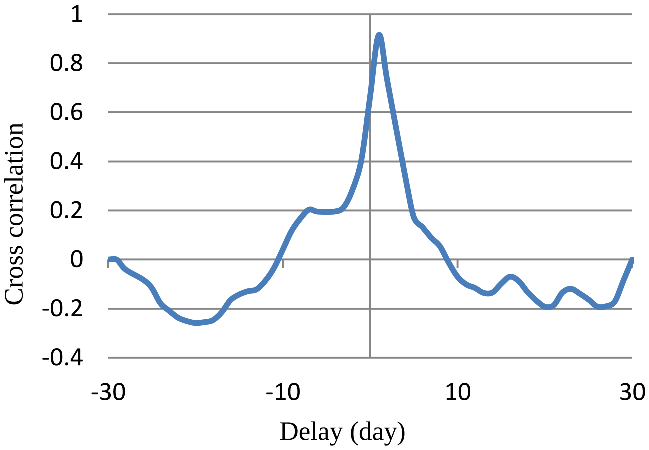

Fig. 3. Cross correlation between Google hit counts and Wikipedia page views of Jared Cohen

III. Results

We calculated the strongest cross correlations for terms in the four categories and recorded them together with their delays. The results are shown in Figure 4 and Figure 5. The horizontal axis represents the delay in days and the vertical axis represents the strongest correlation caused by the according delay. We can see from the figures, that most of the cross correlations lie between 0.2 and 0.8 with delays ranging from -20 days to 20 days.

Table I displays the average and the standard deviation of the strongest cross correlations between search engine hits and Wikipedia article views. We compare the results among terms from different categories. Yahoo! and Bing show a stronger correlation than Google for terms in medicine, science and technology, while Google has a stronger correlation for queries about people. Google, in general, has a larger standard deviation than Bing and Yahoo!.

Terms in the medicine field result in the lowest average cross correlation between the search engine hits and Wikipedia page views. One of the reasons is that, unlike terms in other categories, medical terms often have synonyms. For example, bovine spongiform encephalopathy has a synonym of mad cow disease, which is more commonly known. At the time of writing, search term "bovine spongiform encephalopathy" results in 444,000 Google hits, while query "mad cow disease" returns 1,640,000 Web pages in Google. The former result fails to represent the total number of pages on the Web regarding the disease. One possible solution to this problem is to include the hits of synonyms, which we will discuss in Section IV.

Table II shows the average and the standard deviation of delays that result in the maximum cross correlation. It is interesting to see that Bing/Yahoo! has an average positive delay except for the science category, while Google has negative delays for all categories. Comparing with Bing and Yahoo!, Google has a larger standard deviation.

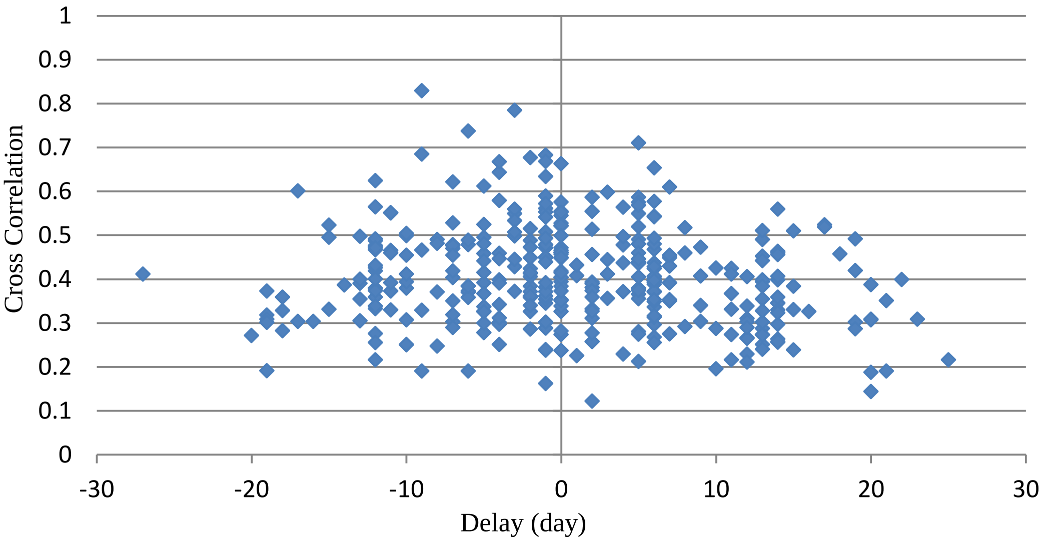

Fig. 4. Maximum cross correlations with delays between Bing/Yahoo!'s hit counts and Wikipedia article traffic of all terms

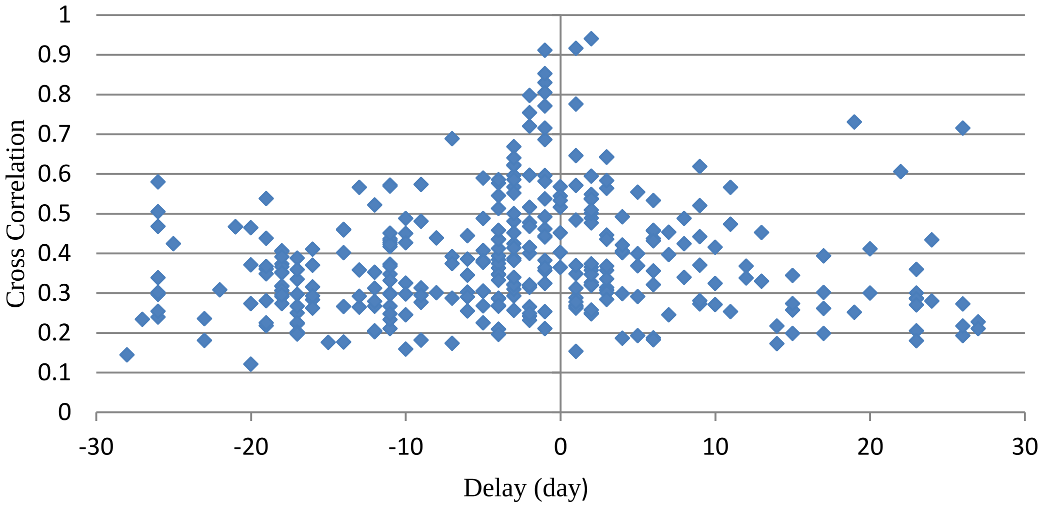

Fig. 5. Maximum cross correlations with delays between Google's hit counts and Wikipedia article traffic of all terms

TABLE I. AVERAGE AND STANDARD DEVIATION OF THE STRONGEST CROSS CORRELATIONS

	Average Cross Correlation (Bing/Yahoo!)	Standard Deviation (Bing/Yahoo!)	Average Cross Correlation (Google)	Standard Deviation (Google)
Medicine	0.37	0.10	0.34	0.11
People	0.41	0.11	0.46	0.17
Science	0.41	0.11	0.35	0.11
Technology	0.44	0.12	0.42	0.17
Overall	0.41	0.11	0.39	0.15

TABLE II. AVERAGE AND STANDARD DEVIATION OF DELAYS

	Average Delay (Bing/Yahoo!)	**Standard Deviation (Bing/Yahoo!)**	**Average Delay (Google)**	**Standard Deviation (Google)**
Medicine	0.04	9.61	-4.19	10.75
People	2.92	9.48	-1.60	10.32
Science	-3.15	9.48	-3.31	13.82
Technology	2.36	7.94	-2.90	11.51
Overall	0.78	9.37	-3.06	11.56

As mentioned in Section II, a positive delay in the correlation represents a delay in the Wikipedia page views, while a negative delay indicates a delay in the number of websites. In other words, Wikipedia article traffic can be predicted by the number of related Web pages, if there is a positive delay. Similarly, the Web hosting trend can be predicted by the growth/decline of the Wikipedia article views, if there is a negative delay. Among all the search terms, Bing and Yahoo! produce 53.3% positive delays and 46.7% negative delays while calculating the cross correlation. On the other hand, 62.7% of the cross correlations with Google were generated by negative delays, as shown in Table III.

TABLE III. DISTRIBUTION OF POSITIVE AND NEGATIVE DELAYS

	Positive Delays	**Negative Delays**
Bing/Yahoo!	53.3%	46.7%
Google	37.3%	62.7%

IV. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a method to analyze the correlation between search engines' hit counts and Wikipedia's article traffic. We collected 400 popular search terms from fields of medicine, people, science and technology. Each term was sent to major search engines, including Google, Bing and Yahoo!, to retrieve the number of related pages on the web. We also extracted the number of views on the term's Wikipedia article. A month of data was collected and used to calculate the cross correlations between the search engine hits and Wikipedia page views. Most cross correlations lie between 0.2 and 0.8, which represent a moderate to strong positive relationship.

We also analyzed the predicting trends between the hits on the Web and the views in Wikipedia. 62% of the cross correlations between Google and Wikipedia were caused by negative delays, which indicates a leading trend in the Wikipedia article traffic in predicting the inclination of Web hosting. As mentioned in Section III, a term with synonyms may cause inaccuracy in calculating the correlation. In the future, we plan to use a synonym search API, such as [13], to include the search engine hit counts returned by synonyms.

ACKNOWLEDGMENT

We thank Rocco Pascale for his contribution to the collection of the search engine hit counts and the Wikipedia article views.

REFERENCES

[1] T. Tian, S.A. Chun, and J. Geller, "A prediction model for Web search hit counts using word frequencies," Journal of Information Science, Sage Publishing Co., vol. 37, issue 5, pp. 462-475, 2011.

[2] L. Yuen, M. Chang, Y.K. Lai, C.K. Poon, "Excalibur: a personalized meta search engine," the 28th Annual International Computer Science Software and Applications Conference, vol. 2, pp. 49–50, 2004.

[3] R.L. Cilibrasi, and P. Vitanyi, "Normalized Web distance and word similarity," in: Handbook of Natural Language Processing. 2nd ed., N. Indurkhya amd F.J. Damerau, Eds. Boca Raton, FL: CRC Press, Taylor and Francis Group, 2010.

[4] D. Milne, and I. H. Witten, "Learning to link with Wikipedia," the 17th ACM Conference on Information and Knowledge Management, pp. 509 - 518, 2008.

[5] S.P. Ponzetto, and M. Strube, "Knowledge derived from Wikipedia for computing semantic relatedness," Journal of Artificial Intelligence Research, vol. 30, pp. 181-212, 2007.

[6] S. Cucerzan, "Large-scale named entity disambiguation based on Wikipedia data," the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, pp. 708–716, June 2007.

[7] M.R. Laurent, T.J. Vickers, "Seeking health information online: does Wikipedia matter?" Journal of the American Medical Informatics Association, vol. 16, issue 4, pp. 471-479, July 2009.

[8] Wikipedia article traffic statistics, http://stats.grok.se/, retrieved on 04/15/2015.

[9] Google's JSON/Atom Custom Search API, https://developers.google. com/ custom-search/json-api/v1/overview, retrieved on 04/15/2015.

[10] Bing Search API, http://datamarket.azure.com/dataset/bing/search, retrieved on 04/15/2015.

[11] D. Milne, and I. H. Witten, "Learning to link with Wikipedia," the 17th ACM Conference on Information and Knowledge Management, pp. 509 - 518, 2008.

[12] M. Andrews, "Searching the Internet," IEEE Software, vol. 29, issue 2, pp. 13- 16, 2012.

[13] R. Bracewell, "Pentagram Notation for Cross Correlation," The Fourier Transform and Its Applications, New York, McGraw-Hill, pp. 46 and 243, 1965.

[14] Stands4 API, http://www.abbreviations.com/api.asp, retrieved 05/17/2015.