INTERNET SEARCH STATISTICS AS A SOURCE OF BUSINESS INTELLIGENCE: SEARCHES ON FORECLOSURE AS AN ESTIMATE OF ACTUAL HOME FORECLOSURES

G. Kent Webb, San Jose State University, webb_k@cob.sjsu.edu

ABSTRACT

Historical search data, describing the volume of searches by topic and region, have recently become freely available. This provides a potentially valuable source of data useful for business intelligence about conditions external to the organization where data is sometimes sparse. As an experiment for a business application, Google searches on the keyword "foreclosure" were correlated with actual U.S. home foreclosures over the past 4 years. The resulting regression analysis shows a very good correlation, indicating that searches on "foreclosure" provide a very accurate estimate of trends in actual U.S. home foreclosures and may provide an early warning system. In a related non-business experiment, Google has recently reported success in showing that searches on the term "flu" track closely with worldwide outbreaks of flu.

Keywords: Search Statistics, Business Intelligence, Google Trends

INTRODUCTION

Bringing external data into the organization creates special challenges to information systems bridge building. Finding useful and cost effective data can be difficult. Transforming the data into information that can help decision makers in turbulent economic times can be critical to the survival and success of the organization [23]. This paper examines a relatively new data source, Google Trends that provides an index of weekly keyword searches over the past several years.

Given the particular importance of the housing market in the current economic environment, an experiment was conducted to see if this search data could be used as business intelligence to identify trends in U.S. home foreclosures. Google provides daily updates of weekly data, making it among the most current data available. Having very current data is often important for business intelligence. The statistical analysis described in this paper shows a very strong correlation between searches on the keyword "foreclosure" and actual U.S. home foreclosures, suggesting value as a source of business intelligence for organizations concerned with that market. More generally, the results suggest that the keyword search data could be a valuable source of external data for a variety of information systems where external data is an important input such as for marketing [28], business forecasting [26], environmental scanning [24, 25], executive information [27], and business intelligence in general [30].

As early as 1997, Pawar and Shara [1] recognized the potential of the internet as a source of external data for business intelligence. They note the characteristic problems associated with bringing external data into the organization's information system: less structured and less controlled.

Several studies discuss the increasing importance of external business information to an organization for strategic decision-making and propose methods for identifying and capturing useful data on the internet [7, 8, 9, 10, 12, 14, 15]. Because the internet has some capacity as a self-organizing system, it is not surprising that the best data mining on the web might be located at some hot spots specifically designed to provide useful sources of data. Google Trends may be one such site.

Researchers at Google have recently demonstrated that search data closely tracks the actual outbreaks of flu as measured by a surveillance program managed by the U.S. Centers for Disease Control (CDC) and Prevention, but that estimates provided by search data allowed accurate estimates of flu conditions one to two weeks faster than the CDC reports [2].

Rech [3] used Google search data provided in the Google Trends service to identify trends in software engineering, although no effort was made to verify the reliability of this data against other data. Webb [4] found that search data could produced similar results to a traditional survey that rated the top green technology investments [5]. Radinsky, Davidovich, and Markovitch have recently demonstrated that they can accurately predict the terms that will show up in

the news up to one week forward using data from Google Trends [5].

Noting the importance of identifying future business trends, Liu [6] develops a model to predict trends in keyword statistics using classic time series models. An informal effort to predict the presidential primary elections is described in an online blog [18]. All of these efforts are an attempt to use the keyword search data to detect changes or patterns in behavior. This might be just the data we need to expand some of our forecasting efforts.

Former Federal Reserve Chairman Alan Greenspan, commenting on the failure of econometric models to predict the level of the problem currently facing the U.S. economy related to the huge increase in foreclosures suggests that the models failed to include information on the "human responses that result in swings between euphoria and fear." Keyword search data may be useful in this regard.

Google Trends

To support bidding on keywords among advertisers using its services, Google has been experimenting for a few years with providing keyword search statistics on the internet site at http://www.google.com/trends. Users can type in a keyword and get data and a graph of the index for the volume of searches over the past several years. For example, typing in the word "foreclosure" creates the following graph in Figure 1. The dramatic increase in searches on foreclosure illustrated in the figure at the beginning of 2008 seems to signal an increase in interest about the topic.

Figure 1: Google Trend Results for "Foreclosure"

As an illustration of the potential usefulness of the keyword search data, the Google Trends site provides a number of examples showing how the search volume for keywords coincides with seasonal patterns in the industry. Searches for "summer camps" [16] increase near the end of the school year when parents may be planning for summer recreation. Searches for the "Internal Revenue Service: increase around April 15, corresponding to the tax deadline [17]. Some mainstream media content providers are watching the Google Trends results to help determine the topics for editorial content [20].

One serious problem that has perhaps restricted the use of the Google Trends service for business and academic research has been some sporadic support in the recent past. At the end of 2006 and into early 2007, the data was not updated; but it has been updated regularly since then. Currently, the data is updated daily.

An indication, however, that the service may continue is the development of an extension of Google Trends called "Insights" that is designed for more detailed analysis, available since about August, 2008 at http://www.google.com/insights/search/#. The service is specifically designed to provide external data for market analysis, but one economist sees it as having potential in "economic forecasting, finance, and sociological studies." [19]. The Insights web page suggests that in terms of seasonality the data can help "anticipate demand for your business so you can budget and plan according."

Although there appears to be no academic research yet citing the use of Google Insights, a search on the terms results in a number of blogs and commentaries expressing enthusiasm for this new research tool, particularly in marketing applications.

RESEARCH METHODOLOGY

The basic research question here is if the Google search data can provide a useful source of business intelligence, specifically for the U.S. housing market in this example.

The research hypothesis to be tested is:

H_1: Internet searches on the keyword "foreclosure" correlate with actual U.S. home foreclosures

Since the search data is often several weeks ahead of any other data, it could serve as an early warning system. This result would suggest that the search data might also be helpful for a variety of other applications as well.

Data Sources

The indexed historical search volume for "foreclosure" that is illustrated in Figure 1 comes

from Google Trends [10]. This data can easily be downloaded into a .csv (Excel compatible) spreadsheet. Actual volume data is not available, but two types of indexed data are available: relative and fixed.

Google does not disclose exactly how the indexes are created but advises that the data may "contain inaccuracies for a number of reasons, including data-sampling issues and a variety of approximations [21]." Both the relative and fixed indexes are scaled so that 1.0 represents the average search traffic during a time period. In the relative mode, the index is set to 1.0 for the beginning of the time period selected by the user. In fixed mode, the index is set to a fixed point, usually January 2004, even if the user specifies another time period. The data used in this study is the fixed index beginning in January 2004.

The following disclaimer also appears on the site: "All results from Google Trends are normalized, which means that we've divided the sets of data by a common variable to cancel out the variable's effect on the data and allow the underlying characteristics of the data sets to be compared."

Data used in this study for total monthly U.S. home foreclosures comes from RealtyTrac [11], a market research firm that releases monthly reports by state and the total U.S. market compiled from government statistics. The actual foreclosure data is two weeks to one month old while the Google data is usually within a week. The more rapid availability of the Google data is one of the important potential advantages.

The Google data is weekly, while the foreclosure data is monthly. For the statistical analysis that follows, the weekly index values were averaged over the month to create a monthly times series of Google search results. This averaging process results in some loss of information and might be an area of improvement for future of data analysis.

An ordinary least squares regression analysis was conducted using the RealtyTrac market research tallies of actual foreclosures each month from January 2005 to December 2009 as the dependent variable. The search index for "foreclosures" is the independent variable. Although the Google Trends data is available back to 2004, the RealtyTrac foreclosure data is only available from January 2005.

The model specification is intended to capture the correlation between the two variables, but also suggests how a short-term prediction model can be set up. As new Google Trends data becomes available, it could be feed into a regression generated equation that would estimate the new foreclosures numbers before they become available from the market research data.

Figure 2 illustrates the relationship between the actual foreclosures and the monthly average of the index of foreclosure searches. At the beginning of the period for which data is available, 1995, the search index dips a bit but begins to accelerate into the fall of 2005 coincident with the first big uptick in actual foreclosures. The search data also coincides with the first large downturn in foreclosures in fall of 2007. Into 2008, both searches on foreclosures and actual foreclosures were on the rise.

RESULTS

As reported in Table 1, the Search Index data provides a surprisingly good estimate of actual U.S. home foreclosures. The p-value is so small that it is presented in scientific notation; the statistical significance is so high that the data was carefully checked to be certain there were no analytical errors. It is significant well above the 0.05 or 0.01 levels. The null hypothesis of no correlation is rejected. The statistics strongly support the research hypothesis that there is a correlation between actual foreclosures and searches on foreclosure.

This result is similar to the result reported by the Google group in tracking flu outbreaks. Since the Google search data is available before the CDC flu data or, in this example, the actual foreclosure data, it seems to provide a reliable early warning system. This can be quite valuable for business intelligence.

Table 1: Regression of U.S. Home Foreclosures Against an Index of Searches on Foreclosure From Google Trends

Dependent Variable: U.S. Home Foreclosures (Monthly)		
R Square = .87		
Variable	*Coefficient*	*p-Value*
Constant	- 23655.1	0.04*
Keyword Searches On Foreclosure	123857.7	8.9E-22**
N = 48		
* Statistically significant at .05 level		
** Statistical significant above the .00001 level		

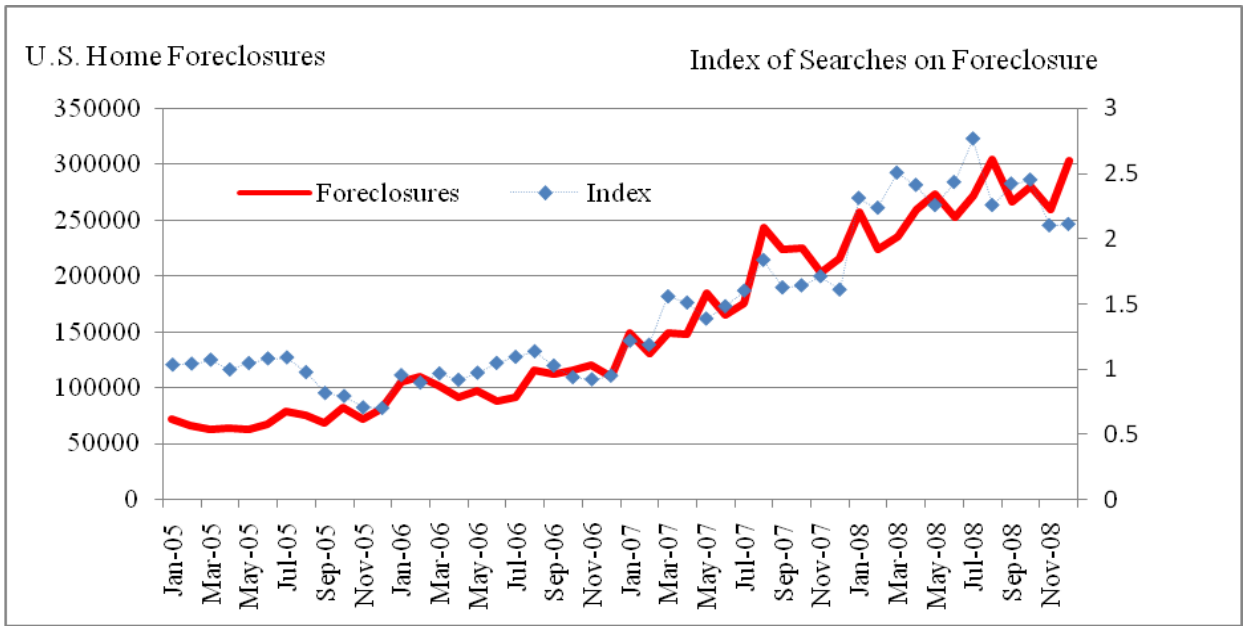

Figure 2. U.S. Monthly Foreclosures (Left Axis) and an Index of Monthly Searches on the Keyword Foreclosure (Right Axis)

CONCLUSIONS

The very strong correlation between searches and foreclosures suggests the Google Trends data could be useful for organizations wanting external data related to the U.S. housing market. One advantage is that the search data is available weekly and sooner than the market research statistics that are currently available to track foreclosures.

More generally, this result suggests that keyword searches can be a useful source of business intelligence that can be captured by an organization at a very low cost. Future research might focus on implementation issues such as a more thorough analysis of the benefits of this data for an organization with a specific interest in the housing market. A more extensive modeling effort to see how the search data can be used to build business-forecasting models is also another logical next step.

REFERENCES

1. Pawar, S., & Sharda, R. (1997). Obtaining business intelligence on the internet. *Long Range Planning*, 30(1).110-121

2. Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., and Brilliant, L. (2008). Detecting influenza epidemics using search engine query data. *Nature*. (doi:10.1038/nature07634). Published online 19 November 2008. Available at http://www.nature.com/nature/journal/vaop/ncurrent/full/nature07634.html

3. Rech, Jorg (2007). Discovering trends in software Engineering with Google Trend. *ACM SIGSOFT Engineering Notes*. March. 32(2), 1 – 2.

4. Webb, G.K. (2007). Analysis of pages and metrics related to global environmental management. *Issues in Information Systems*, 9(2), 111-116.

5. Radinsky, K., Davidovich., S. & Markovitch, S., (2008). Predicting the news of tomorrow using patterns in web search. *Proceedings of 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Agent Technology*, 363- 367.

6. Liu, N., Yan, J., Yan, S., Fan, W., & Chen, Z. (2008). Web query prediction by unifying model. *Proceedings of the 2008 IEEE International Conference on Data Mining Workshops*. 436-441

7. Rouibah, K & Ould-ali, S. (2002). Puzzle: a concept and prototype for linking business intelligence to business strategy. *The Journal of Strategic Information Systems*,11(2). 133-152 (doi:10.1016/S0963-8786(02)00005-7)

8. Heinrichs, J.H. & Lim, J. (2003). Integrating web-based data mining tools with business models for knowledge management. *Decision Support Systems*, 35(1), 103-112 (doi:10.10`6/S0167-9236(02)00098-2)

9. Wood, E. (2001). Marketing information systems in tourism and hospitality small and medium-sized enterprises. *International Journal of Tourism Research* 3, 283-299 (DOI: 10.2003/jtr.315)

10. Google Trends, available at: http://www.google.com/trends.

11. Realtytrac, available by subscription at: http://www.realtytrac.com

12. Azvine, B., Cui, Z., & Nauck, D.D. (2005) Towars real-time business intelligence. *BT Technology Journal*, 23(3). 214-225. (DOI 10.1007/s10550-005-0043-0)

14. Marin, J. & Poulter, A. (2004). Dissemination of competitive intelligence. *Journal of Information Science* 30(1), 165-180 (DOI 10.1177/0165551504042806)

15. Teo, S.H. & Choo, W.Y., (2001) Assessing the impact of using the internet for competitive intelligence. *Information & Management* 39(1), 67-83. (doi:10.1016/7206(01)00080-5)

16. Google available at: http://www.google.com/trends?q=summer+camps

17. Google, available at: http://www.google.com/trends?q=IRS

18. Guiffrida, M. Attempt to predict primary/caucus results using Google Trends. Available online at: http://michaelg.us/freakon/googletrends.php

19. Helft, Miguel (2008). Google's new tool is meant for marketers, *New York Times*, August 8, Business Section. Available online at: http://www.nytimes.com/2008/08/06/business/media/06adco.html?_r=1&ref=business.

20. Arrington, Michael (2008). Some big sites are using Google Trends to Direct Editorial. TechCrunch, October 9. Available online at: http://www.techcrunch.com/2008/10/09/some-big-sites-are-using-google-trends-to-direct-editorial/

21. Google, available online at:

http://www.google.com/intl/en/trends/about.html#1

22. Greenspan, Alan (2008). We will never have a perfect model of risk. *Financial Times*, March 16.

23. Martinsons, M.G. (1994). A strategic vision for managing business intelligence. *Information Strategy*, Spring. 17-33.

24. Costa, J (1995). An empirically-based review of the concept of environmental scanning. *International Journal of Contemporary Hospitality Management*. 7(7), 4-10.

25. Frolick, Mark N. (1997) Using EISs for environmental scanning. Information Systems Management. 14(1) 35-40.

26. Heinrichs, John H. (2002) Integrating web-based data mining tools with business models for knowledge management. *Decision Support Systems*. 35(1). 103-112.

27. Xu, X.M.; Lehaney, B; Clarke, S.; Duan, Y (2003). Some Uk and USA comparisons of executive information systems in practice and theory. *Journal of End User Computing*. 15(1) 1-19.

28. Buttery, A. & Tamaschke, R. (1997). Marketing decision support systems and Australian businesses: A Queensland case study and implications towards 200. Journal of Management & Organization. 3(1). 51-58.

29. Felden, C. & Chamoni, P. (2003). Web framing and data warehousing for energy tradefloors. IEEE/WIC International Conference on Web Intelligence. 642 – 648.

30. Watson, J.J., Wixom, B.H. (2007). The current state of business intelligence. Computer, 40(9). 96-99.

31. Gangadharan, G.R. & Swami, S.N. (2004). Business intelligence systems: design and implementation. Information Technology Interfaces. 1, 139-144