Crowd-Squared: Amplifying the Predictive Power of Large-Scale Crowd-Based Data

Erik Brynjolfsson
Sloan School of Management,
Massachusetts Institute of Technology
erikb@mit.edu

Tomer Geva
Recanati Business School,
Tel-Aviv University
tgeva@tau.ac.il

Shachar Reichman
Sloan School of Management,
Massachusetts Institute of Technology
shachar@mit.edu

Abstract

The analysis of large-scale data generated by the crowd has recently attracted extensive interest of marketing scholars and practitioners. Combined with recent advances in computer science and statistics, these data provide a myriad of opportunities for monitoring and modeling customers' intentions, preferences, and opinions. Nevertheless, a crucial step in any "Big Data" analysis is identifying the relevant data items that need to be processed or modeled. Interestingly, this important step has received limited attention in previous research and has been typically addressed by ad-hoc approaches.

In this paper, we offer a novel crowd-based method to address this data selection problem. We label the method “Crowd-Squared,” as it leverages crowds to identify the most relevant elements in crowd-generated data.

To implement this method we developed an online word association game that taps into peoples' "thought collection" process when thinking about a focal term. We empirically tested our approach by comparing its performance to previous studies in three domains that have been used as test-beds for prediction: flu epidemics, the housing market, and unemployment. Our findings demonstrate the effectiveness of this method in providing accurate results that are equivalent or superior to previously used term-selection methods.

Introduction

In recent years, there has been growing interest in the opportunities of using and analyzing large scale data generated by crowds. Such data allow analysts to conduct real-time, large-scale monitoring of customers' intentions, preferences and opinions, and to model and explain economic phenomena.

This abundance of data also creates significant challenges for data collection and processing. Perhaps one of the most important challenges to emerge is how to determine which data items should be selected for modeling a phenomenon of interest.[1] Specifically, given the vast available data (for example, all possible search engine queries), it is essential to select the specific data items (continuing with the previous example, specific query terms) that are relevant for a given modeling or prediction task.

As illustrated in Figure 1, which outlines the prediction process using large-scale online crowd-generated data, data selection step (step 2) is a critical stage that bridges between the data generated by crowds and the actual inputs for modeling. Notably, this selection stage involves the first decision made by the modeler.

Interestingly, this important step received only limited attention in previous literature and has been commonly performed using three main approaches: (1) intuition and prior knowledge, (2) algorithmic classification methods, and (3) a comprehensive scan of the data.

[1] It is important to note that the data selection problem to which we refer is substantially different from the well-known feature selection problem (Liu & Motoda 1998). The feature selection problem concerns the challenge of selecting a subset of informative variables (A) out of a larger, initial set of features (B) that are already available to the researcher. In contrast, the data selection problem addresses the question of which data, out of all possible data that potentially could be collected -- should actually be collected by the researcher.

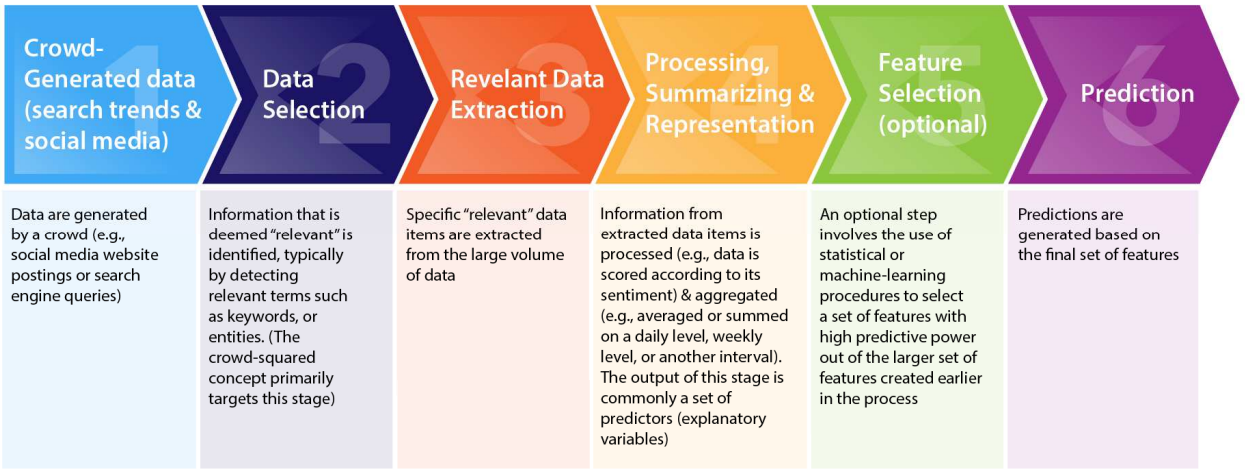


Figure 1 – The Prediction Process Using Online Crowd-Generated Data

Using the first approach, the research applies human intuition and prior knowledge to identify online data pertaining to a certain item (e.g., flu). This is commonly performed by choosing potentially relevant keywords (e.g., "flu", "influenza") to identify relevant subset of data.

The second approach uses automated methods to classify data into predefined categories (such as Google Trends' internal category classifier).[2] These classifiers commonly focus on detecting items that pertain to a pre-determined category (e.g., search queries that relate to the "Ford" car model). However these classifiers do not take into account the context of the prediction task and therefore might be unsuitable for detecting data items with "indirect" relevance for a given prediction task (For example, Wu and Brynjolfsson, 2009, found that search trends related to new home purchases were shown to be indicative of future home appliance sales).

[2] Google Trends is a publicly available product that aggregates billions of search queries and provides information about the relative volume of different search terms. http://www.google.com/trends/

The third approach uses a comprehensive scan of all available data to select the terms that are most strongly correlated with the focal phenomenon. Previous studies using this approach commonly resorted to the use of proprietary data (Ginsberg et al, 2008). Moreover, such analyses require extensive computational power to detect correlated data to the phenomenon of interest.

In this paper, we offer a novel crowd-based approach to this data selection problem. We label it "crowd-squared," as it leverages crowds to amplify the predictive capacity of crowd-generated data. We apply a simple, inexpensive implementation to demonstrate the predictive capacity of the "crowd-squared" approach in comparison to data selection methods used in previous studies. Specifically, our demonstration of the predictive capacity of the "crowd-squared" concept utilizes Google Trends[3] and a crowdsourcing environment designated "game."

To capture people's ideas of potentially relevant terms, we employed an online word association game in which consumers are asked to provide terms that come to mind when they view a specific word or phrase. Given a specific phrase, word association techniques provide a relative index of the accessibility of related words in the memory. With the wide use of the Internet as a primary form of external or transactive memory (Sparrow et al., 2011), we expect this association technique to simulate the same keyword-generating process that occurs when one uses a search engine (Nelson et al., 2004).

After using this method to identify the most popular terms, we collected corresponding search trend data, and generated predictions in three different domains: influenza

[3] Researchers have previously utilized Google trends to make accurate predictions of a wide variety of future events, including products sales, claims for unemployment, and epidemic outbreaks.

epidemics, unemployment claims, and housing indexes. We then compared our results with a well-known benchmark model in each domain. We found that the use of the crowd-squared method was highly effective. Our results suggest that the integration of crowd-selected search terms with aggregated data from search engines performs as well as or even outperforms these benchmarks – and does so at a very low modeling cost. Additional advantages of our methodology are improved understandability and finer-grained analysis capabilities compared to several benchmark methods.

Related Literature

The availability of search data, web activity data, and other sources of information, along with developments in analytic tools, have dramatically increased our ability to obtain accurate data on millions of economic decisions, as well as on individuals' intentions to make transactions (McAfee and Brynjolfsson, 2012). In the past decade, the use of large-scale data generated by crowds to explain and predict various economic outcomes has become commonplace in scientific research.

Specifically, marketing and other management literature commonly report the utilization of information from social media websites, such as online reviews, discussion forums and blogs in which crowds can communicate product information and WOM (word of mouth) to each other. Early work in this domain includes studies such as Godes and Mayzlin (2004), Chevalier and Mayzlin (2006), and Liu (2006). Since then this research field gained significant momentum which resulted in a large number of scientific studies.

At the same time, search engine logs or search trends, aggregating large volume of

crowd-generated search queries, have received significant attention for their utility in detecting and predicting a variety of economic outcomes. Search volume data have been shown to provide useful predictions in a wide range of domains, from epidemic outbreaks (Ginsberg et al., 2008), through movie box office sales and music billboard rankings (Goel et al., 2010), to automotive sales (Choi and Varian, 2012; Du and Kamakura, 2012; Geva et al., 2013), home sales (Choi and Varian, 2012; Wu and Brynjolfsson, 2009), unemployment claims (Choi and Varian, 2012), and private consumption (Vosen and Schimdt, 2011).

However, even with the availability of powerful aggregation tools and advanced text-processing tools (e.g., Netzer et al. 2012), predictive modeling using crowd-based data still depends on a critical aspect—which data are selected for modeling the phenomenon of interest.

Selecting specific relevant terms is a challenging task. Online data items that refer to a specific phenomenon might do so using any number of terms (e.g., influenza may also be referred to as flu or cold). In straightforward cases, the keywords associated with an item of interest may include sub-items from known ontologies (e.g., online mentions of various Chevrolet models such as Aveo or Camaro are likely to be indicative of interest in the Chevrolet brand). In other cases, terms indicative of or correlated with a certain item of interest may not include a direct reference to the item of interest or its sub-items. For example, online searches for "inexpensive cars" may also contain valuable predictive information regarding consumer interest in certain brands such as Chevrolet. In other cases, the relevance of a keyword to a phenomenon of interest may be even less direct (e.g., home purchases were shown to be indicative of future home appliance sales; Wu and Brynjolfsson, 2009).

Term selection in the specific context of social media data involves additional complications. Unlike search queries that typically include a limited number of words, social media data appear in more a complex textual format that includes sentences and paragraphs, and may include complex structures such as anaphoric references. Thus, the modeler using social media data must carry out extensive text processing as well as data aggregation in order to generate predictors (explanatory variables). Therefore, choosing the "wrong" set of initial terms may entail an expensive process of rework to re-identify relevant data items, and re-process and aggregate the information before it can be re-incorporated into the prediction model. (Going back from stage 6 to stage 2 in Figure 1).

In practice, most previous studies using social media to predict or explain economic outcomes either: (a) utilized straightforward terms to identify relevant data items (e.g., Dhar and Chang, 2009, used music album titles and Rui et al., 2013, used movie name mentions on Twitter to identify WOM relating to the mentioned products); or (b) focused only on specific websites in which an item of interest is clearly identified: For example, Chintagunta et al. (2011), Dellarocas et al. (2007), Duan et al. (2008), Liu (2006) and others used user review data from the Yahoo! website, in which reviews for a given movie are posted on a webpage dedicated to that movie. Similarly, Dewan and Ramaprasad (2012) used identifiable reviews for songs on the Amazon website, and Chevalier and Mayzlin (2006) used designated book reviews on Amazon and BarnesAndNoble.com. While focusing exclusively on clearly identifiable data is suitable for various research goals, it clearly limits the possibility of using additional data from many other websites in which information is less directly linked to a given item of

interest. Furthermore, in various domains, clearly identifiable data may not be available or may be limited in scope.

In the case of search trend data, in addition to the use of simple terms suggested by researchers (e.g., D'Amuri and Marcucci 2012), several more advanced approaches have also been utilized to address the data selection problem. The first approach relies on using a "black box" automated category classifier. This approach utilizes the category assignment provided by a category classifier available on the Google Trends website. This classifier can categorize search queries into several hundred predefined categories and sub-categories, and has been utilized in various studies such as Choi and Varian (2012), Wu and Brynjolfsson (2009), and Vosen and Schimdt (2011). While such classifiers can encompass multiple relevant search terms, they are effectively "black boxes" to users not affiliated with the Google classifier's developer. As a result it is difficult to gauge their accuracy or coverage.[4] Additionally, it is possible that the classifier's rules were determined (or examples were provided for a supervised learning-based method) in a "one-person guessing game." Furthermore, in Google's popular classifier, the predetermined categories are applicable only to a set of popular items, but exclude many potential items of interest (e.g., there is a category for the Ford automotive brand, but there is no category for the Ford Focus model).[5]

Another approach was adopted by Ginsberg et al. (2008), who constructed an early detection system for influenza epidemics. The researchers used Google's internal data

---

[4] Another "black box" source for keywords is Google AdWords, which is commonly used to recommend relevant search terms for advertising purposes, however it has also been used for search term selection in predictive studies such as Du and Kamakura (2012).

[5] Google Trends also allows users to specify a keyword within a Google category. For example, a search for the term "Argo" under the "movie" category will return search queries related to movies that specifically include the word "Argo." Various studies (e.g., Geva et al., 2013; Seebach et al., 2011) used both Google category classifier and individually selected relevant keywords.

concerning the 50 million most popular search terms and performed a comprehensive scan over these data to select the terms that correlated most strongly with actual influenza data. However, it is impossible to reproduce this methodology with the search trend data that Google provides to external users (on the Google Trends website), due to strict limit on the number of terms that can be extracted from Google Trends (several hundred per day). In addition, this kind of analysis required expertise and computational power to create the correlation matrix for the phenomenon of interest.

Another study that used proprietary information was conducted by Goel et al. (2010). This study reported various methodological aspects of using search trends data. To demonstrate these aspects, they performed tasks such as predicting movie revenues, music billboard rankings, and video game sales. Their relevant keyword identification methodology relied on the identification of search queries using predefined relevant webpages (e.g., in IMDB) that were returned by the Yahoo search engine when these search terms were entered as input. While the authors (who were affiliated with Yahoo) obtained good results using this methodology, it is virtually impossible to replicate their method using publicly available data, as this entails an exhaustive check of all possible search terms that may return a set of predefined links.

In the research reported in this paper, we use a crowdsourcing technique to identify relevant information in large-scale crowd-based data. We use search trends, which aggregate a large number of search queries, as our test-bed. One important aspect of search queries, which makes search trends a suitable test-bed for our methodology evaluation, is that search query texts are commonly brief and focused. As a result, the use of search trends allows a direct application of our proposed method compared to the

use of various intervening procedures that are required to extract terms from complex texts.

The fundamental idea behind prediction based on search trends data is that these data reflect cumulative actions performed by people over time and, as a result, capture longitudinal changes in behavior. We propose using the crowd to better understand how individuals decide on the keywords they use in their search queries. As search behavior can be used to reveal consumers' intentions (Moe and Fader, 2004), improved understanding of the keyword generation process could improve classification of search patterns of different consumption activities.

Crowdsourcing is the act of harnessing a distributed network of individuals to solve a problem or perform a function that was once performed by employees (Brabham, 2008; Howe, 2006). In recent years, the use of crowdsourcing has grown dramatically in many fields and tasks such as capturing new product ideas and innovations (Bayus, 2013), generating accurate image tags (Von Ahn, 2006), improving image search (Yan et al., 2010), and even solving scientific problems (Lakhani et al., 2007). Crowdsourcing has also been used to aid in processing social media data. For example, Archak et al. (2011) used crowdsourcing to extract product features. Overall, the benefits of crowdsourcing stem from its scale and from the diversity of user backgrounds, levels of expertise, and other demographics, coupled with its low costs. We follow this stream of research and leverage the crowd to generate relevant keywords for prediction and early detection of events with search volume data.

One of the challenges of crowdsourcing is how to engage the crowd in a meaningful and productive manner (Boudreau et al. 2013). As noted by Von Ahn (2006), an online game environment is an effective setting for capturing crowd knowledge and may be used to

elicit reliable information without any supplementary verification of users' answers. Furthermore, as shown by Snow et al. (2008), aggregating results for the same task from multiple non-expert individuals can generate results at the same level as those created by experts.

In this paper, we demonstrate a crowd-squared-based approach using a crowdsourcing game environment that utilizes a word association game to capture people's ideas of focal phrases. We aggregated the resulting terms, collected search data for each of the most frequently mentioned terms, and included them in the prediction model.

Methodology and Evaluation

We studied how a crowd-based word association game can improve the generation of useful search terms, thereby improving trend predictions. We used the Amazon Mechanical Turk platform, an online marketplace for tasks that require human intelligence (or tasks that are easily answered by a human but require large computation costs to be solved algorithmically). Workers (known as Turkers) are paid small amounts of money to complete small tasks (called HITs – Human Intelligence Tasks). The platform allows randomization of task assignments to multiple Turkers and provides control over task completion. In total, 300 Turkers participated in our experiments.

Word association

To demonstrate the capabilities of the crowd-squared concept, we introduced a technique to use human workers to help us identify relevant keywords in a game-like environment. Specifically, we implemented a word association game (also known as free association) where workers were asked to provide related phrases.

Word association is a task that requires participants to spontaneously provide a word or a phrase that is related to a presented word (known as the cue). Word association taps into one's lexical knowledge, which is based on real-world experience (Nelson et al., 2004) and has been shown to be important in predicting cued recall (Nelson et al., 1998). This task is used in everyday activities as a mean for "collecting thoughts" (Nelson et al., 2000).

Word association provides an index of the probability that words are related to the cue term. This information was found to be consistent across different people in the same recall culture (Nelson et al., 1998). In the context of web searches, as people use search engines as a kind of external or transactive memory, word association can be used to determine effective search queries (Sparrow, 2011). With its consistent representations of the associated terms, these terms may reflect broader search patterns and therefore assist in measuring current events and predicting future activities.

Another benefit of the word association technique is that it provides a power law distribution of term associations; Most associations relate to proximal terms, and a few associations connect to more distant terms. This technique allows us to capture terms that are less correlated with each other; thus, they may have more explanatory power when combined with search data.

Keyword association – game design

We designed an online word association website specifically developed for this study. The website contains a single page with brief instructions and one phrase (the cue term). Participants enter their associated terms in five text boxes displayed on the screen (an illustration of this game is presented in Figure 2). The appearance of the website was

planned to simulate a common game environment; Participants were not informed of the purpose of the game or how their terms would be used after the game.

Three hundred participants played the game using the Amazon Mechanical Turk platform. Each participant (Turker) was given a single cue phrase and was asked to provide five terms or phrases that come to mind when seeing this phrase. Each Turker was paid 5 cents ($0.05) for completing the game. The average duration of a game was 46 seconds (including completion of three demographic items).

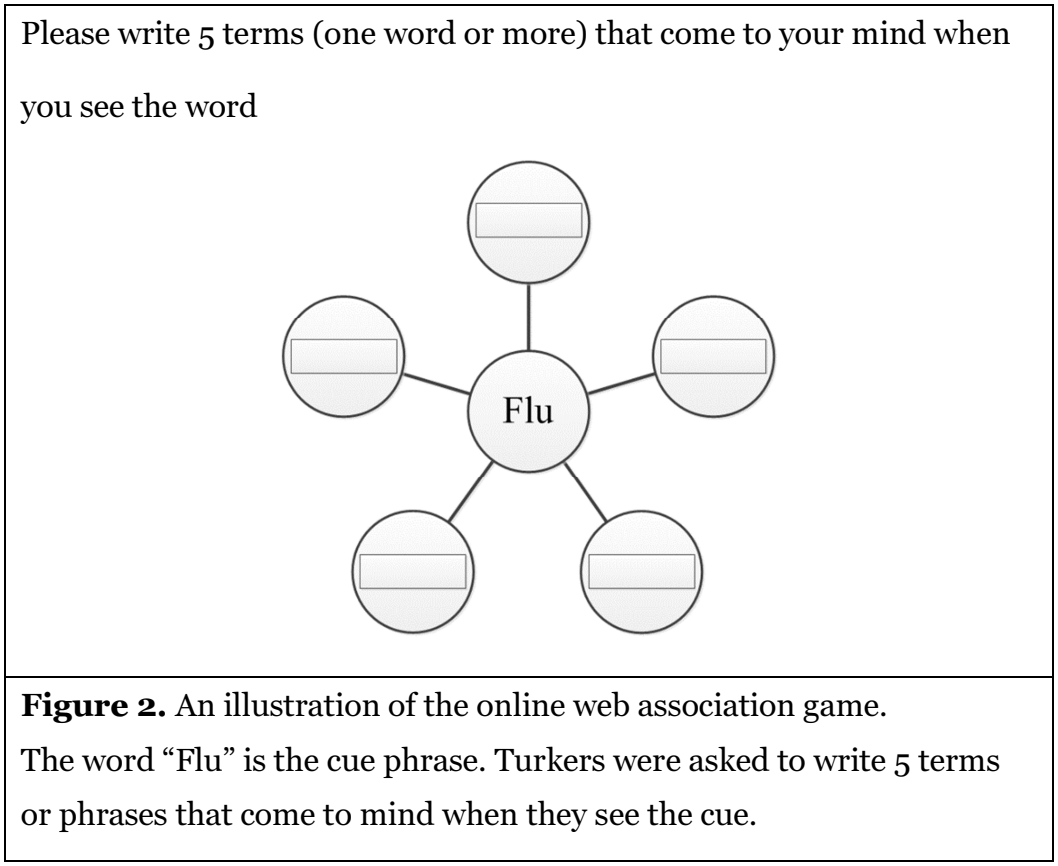


Figure 2. An illustration of the online web association game.
The word "Flu" is the cue phrase. Turkers were asked to write 5 terms or phrases that come to mind when they see the cue.

We aggregated game results and generated a list of the top 10 terms associated with each cue phrase (Appendix A includes the top 10 terms for each cue). We used this set of terms as the list of relevant query terms assumed to accurately reflect actual search

queries. For each term, we collected its search query volume over time and included the search data in the forecasting method.

Evaluation

To validate the effectiveness of the crowd-squared approach, we applied our proposed methodology to similar data and prediction tasks reported in three different domains. We replicated tasks reported in three well-known related studies: Ginsberg et al. (2008) in the influenza outbreaks detection, Wu and Brynjolfsson (2009) in real estate market predictions, and Choi and Varian (2012) in predictions of unemployment levels.

To allow an impartial comparison, we intentionally constrained our analysis to the precise prediction model specifications, performance measures, training data, and validation methodologies specified in each of these studies. The only difference was the data selection methodology. We compared our prediction results with the prediction results reported in each paper and with a baseline model when one was used in the original comparison. If our suggested crowd-squared concept is useful, we expect it to obtain predictive accuracy that is at least as good as the predictive accuracy reported in these studies.

Influenza epidemics

The first dataset that we used to validate our methodology is flu outbreak data from the U.S. Center for Disease Control (CDC). This type of data was used by Ginsberg et al (2008) for constructing an early detection system for influenza epidemics. Specifically, the dependent variable in their study was the weekly ILI (Influenza-Like Illness) factor reported by the CDC. To select the search terms to be included in the prediction model,

the researchers used Google's internal data concerning the 50 million most popular search terms, from which they selected the "top n" terms by calculating individual term correlation with the dependent variable. Subsequently, they used the selected terms to fit a linear model used to generate predictions. Their method was highly successful for this application, achieving an out-of-sample mean correlation of 0.97 across U.S. regions. Nevertheless, it is impossible to use a similar methodology without access to Google's proprietary data since Google does not allow external access to search trend data for more than several hundred search terms a day.

In this study, we used U.S. national-level data from the period between Jan 2005 and the week commencing on March 11, 2007.[6] We validated our modeling using out-of-sample data from March 18, 2007 to May 11, 2008; This is the same out-of-sample validation period used by Ginsberg et al. (2008).

Using the word association setting described above, we asked 100 Turkers (62% female, average age 31.8) to play an online game where the task description was "Please write 5 terms that come to mind when seeing the word *'Flu'*." (see Appendix A for the top 10 list of associated words generated by the Turkers)

The resulting set of different associated phrases was very large. Nevertheless, the use of any single phrase may not represent a common form of thinking but only one's unique thinking that does not reflect other players' search patterns. As shown by Snow et al. (2008), an aggregation of results from multiple individuals can generate results of a high quality. We therefore restricted the analysis to include only the top 10 most popular association phrases.

---

[6] We excluded data from 2003 since Google Trends provides data only from 2004.

For each phrase, we collected the weekly search index from Google Trends. This search index is the share of searches at time t (typically week or month) relative to the total search volume across the time period. We limited our results to queries in the United States to match the predicted variable – flu outbreak in the U.S..

Specifically, we used the following prediction model:

$$ILI(t) = \alpha + \sum_i \beta_i AssociatedTerm_i\,(t) + \varepsilon_i \quad (1)$$

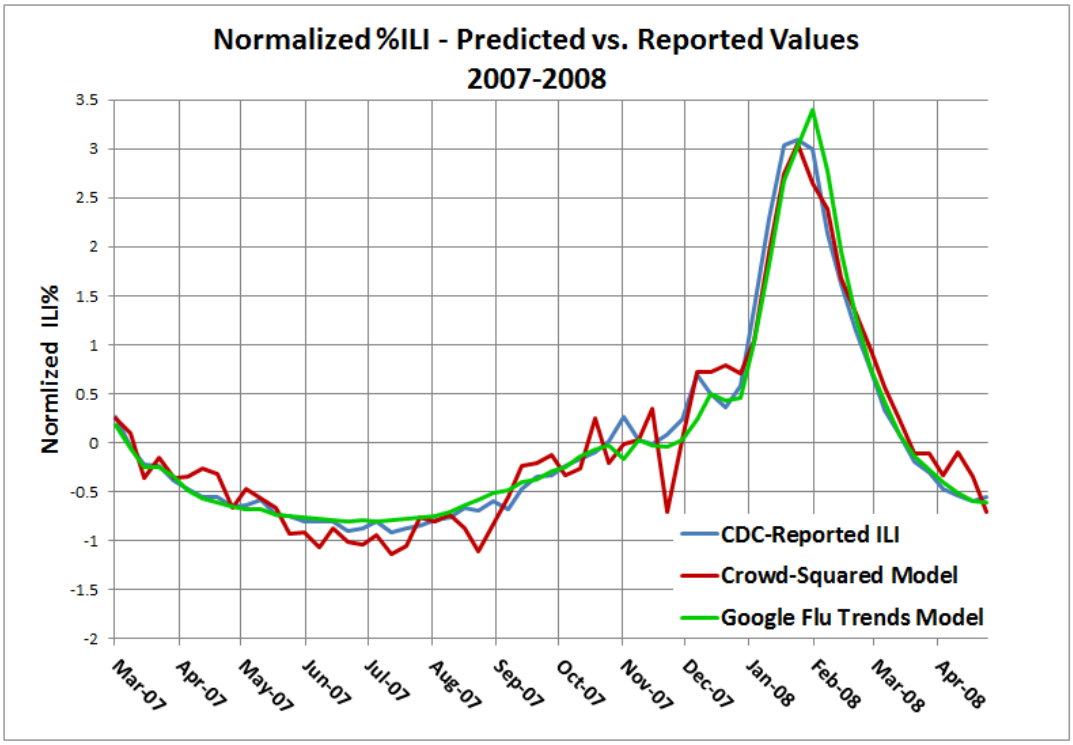
Where *ILI(t)* is the percentage of Influenza-Like Illness at time t as reported by the Center for Disease Control and Prevention (CDC); $AssociatedTerm_i(t)$ is the search trend value at time t for the association-based term i (i=1..10) in the aggregated results of the word association game for influenza.

We first compared the results of our model for the same time period reported in their paper. The training set included 167 weeks from 2004 to 2007. We validated our model on independent out-of-sample data from March 18, 2007 to May 11, 2008. Our prediction results achieved a similar level of out-of-sample correlation (0.973) in predicting ILI (compared to 0.97 in Ginsberg et al., 2008). With seemingly similar results, it is important to point out the huge difference in the volume of data that was included in each model. First, Ginsberg et al. (2008) used 50 million different search terms and 450 million different models to generate the final model that included 45 search term queries. The computation involved in this process employed hundreds of machines using a distributed computing framework. Our method is based on 100 online users; Each played a game for less than one minute. With only the top 10 terms, we generated a single model.

For robustness, we extended our predictions and validated our model using the most recent available influenza data from the first week of April 2012 to the last week of March 2013. We compared our results, based on a prediction model whose most recent training data date from 2007, with flu trend early detection data provided by Google Flu Trends website.[7] This website provides flu outbreak detection on an ongoing basis, using the methodology suggested by Ginsberg et al. (2008). Here, our results show a significant improvement in correlation level, 0.962 compared to 0.951 of the Google Flu Trends results. Figure 3 shows a comparison of our model's predictions with actual reported ILI data from the CDC over the two time periods described above. Looking at the 2012-2013 period, and specifically December 2012 to February 2013, our model generated predictions that better matched the actual influenza outbreak duration compared to the Google Flu Trends model.

To summarize, these results suggest that, with considerably less computation power and with a smaller set of initial candidate search query terms, crowd-squared-based search terms generate equivalent or better results than a significantly more computationally expensive term-selection technique reported in a previous paper.

[7] http://www.google.org/flutrends/us/data.txt.



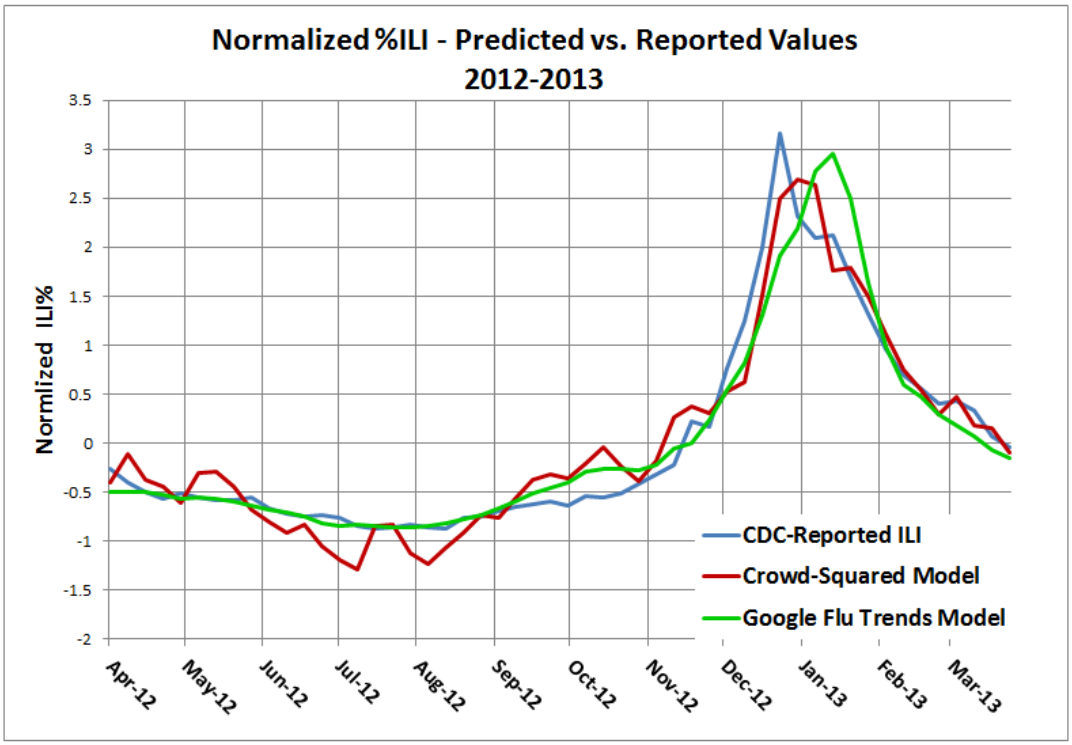


Figure 3. A comparison of the crowd-square model predictions with actual reported ILI and Ginsberg et al. (2008)/Google Flu Trends, over two separated periods: 2007-2008 and 2012-2013.

Housing indicators

The real estate market is traditionally used as a good indicator of a country's economy. Housing activities reflect individuals' financial situations and influence the country's economic growth by generating or eliminating real estate jobs and services. Hence, predictions of real estate indices have become a common and important tool for policymakers and industries that rely on these activities.

This type of data was used by Wu and Brynjolfsson (2009) for predictions of the real estate market and its complementary businesses (such as home appliances). The main predicted variable they used was the volume of housing sales in the U.S.[8] from the fourth quarter of 2007 to the second quarter of 2009. Wu and Brynjolfsson (2009) utilized two hand-picked search term categories for inclusion in the prediction model. Specifically, they used two predefined search categories available from Google's "black box" category classifier: "Real Estate" and "Real Estate Agents." Subsequently, they used the search trend index in the prediction model. Wu and Brynjolfsson used a seasonal autoregressive model and performed an in-sample evaluation of their model using Adjusted R^2. They compared their model to a baseline model presented in equation (2).

Real Estate Indicator models:

$$HomeSales_j(t) = \alpha + \delta_1 HomeSales_j(t-1) + \delta_2 HPI_j(t-1) + \sum S_j + \sum T_j + \varepsilon_j \quad (2)$$

$$HomeSales_j(t) = \alpha + \delta_1 HomeSales_j(t-1) + \delta_2 HPI_j(t-1) + \quad (3)$$

$$+ \sum_i \beta_i AssociatedTerm_i(t) + \sum S_j + \sum T_j + \varepsilon_{ij}$$

[8] Provided by the National Association of Realtors – http://www.realtor.org/research-and-statistics/housing-statistics.

Where $HomeSales_j(t)$ is the volume of homes sales in state j at time t, as reported by the National Association of Realtors; $HPI_j(t-1)$ is the house price index of state j at time $t-1$, as reported by the Federal Housing Finance Agency; and $AssociatedTerm_i(t)$ is the search trends value at time t for the association-based term i (i=1..10) in the aggregated results of the word association game for real estate; S_j is a state-level fixed effect; T_j is a quarterly dummy variable.

We followed Wu and Brynjolfsson's (2009) forecasting methodology and used an autoregressive model presented in equation 3. Similar to the influenza epidemic predictions, we asked 100 participants (53% female, average age 30.6) to play a word association game where the task description was "Please write 5 terms that come to mind when seeing the phrase '*Buying a House*'." (see Appendix A for a list of the top 10 terms associated by participants)

The baseline model (equation 2 above) reported by Wu and Brynjolfsson displayed a good fit with an Adjusted R^2 of 0.973. Our model resulted in an Adjusted R^2 of 0.9882, higher than the highest reported results in their predictions models (0.984). This result provides another demonstration that even a simple crowd-squared method can outperform hand-picked category selection used in conjunction with an automated classifier.

Initial claims for unemployment benefits

The third set of data involves early estimation of the volume of initial claims for unemployment benefits. This economic index is published by the U.S. Department of Labor each Thursday, for the previous (Sunday–Saturday) week and is considered an important measure of the state of the U.S. economy.[9]

Early estimations of initial claims for unemployment using search trend data have been reported by Choi and Varian (2012). Nevertheless, they also report that a simple baseline model, presented in equation (4), performs very well, to the point that linear regression estimation results seem to indicate a random walk (with a drift) behavior.

$$UIC(t) = \alpha + \delta_1 UIC_j(t-1) \qquad (4)$$

Where *UIC(t)* is the logarithm of the seasonally adjusted volume of initial claims for unemployment for week *t*.

Choi and Varian (2012) developed a prediction model that incorporates both baseline information (seasonally adjusted initial claims for the previous week) as well as (seasonally adjusted) search trends for the current week based on Google's predefined categories of "Jobs" and "Welfare…Unemployment," identified by Google's automated category classifier. They evaluated this model out-of-sample using a one-week-ahead rolling prediction (that is, using the data up until week (*t-1*) to train the model and measure its performance over week (*t*)), in the period from January 2004 to July 2011.

While their model was able to generate relatively accurate predictions of economic turning points, their overall results, measured by Mean Absolute Error (MAE), was

[9] Historical data is available at http://www.ows.doleta.gov/unemploy/claims.asp.

3.68%, whereas the MAE for the strong baseline model was 3.37%. This result suggests that the search trend data, based on the predefined categories, may have contained (mostly) overlapping information with the information contained in the previous week's claims data, in addition to some noise that may have reduced out-of-sample predictive accuracy.

We asked 100 participants (54% female, average age 32.8) to play a word association game where the task description was "Please write 5 terms that come to mind when seeing the phrase '*Unemployment*'." (see Appendix A for a list of the top 10 terms associated by participants).

We used this list of top 10 associated trends and reran a simple linear regression model as detailed in equation (5).

$$UIC(t) = \alpha + \delta_1 UIC_j(t-1) + \sum_i \beta_i AssociatedTerm_i\,(t) \qquad (5)$$

Where *UIC(t)* is the logarithm of the seasonally adjusted volume of initial claims for unemployment for week t and $AssociatedTerm_i(t)$ is the search trends value for the association-based term i (i=1..10).[10]

We applied this model using a similar one-step-ahead prediction model and a similar time period as in Choi and Varian 2012 (see Figure 4 for a comparison of the prediction model and actual unemployment claims data). Our prediction model obtained an out-of-sample MAE value of 3.42%. While this value is not as good as the MAE value for the competent baseline model (3.37%), our predictive accuracy was superior to the MAE value reported by Choi and Varian 2012 (3.68%). This suggests that the association-

[10] Associated terms as seasonally adjusted by subtracting the week average value for each term.

based search terms contained less noise than the search volume identified by a hand-picked category combined with Google's automated classifier.

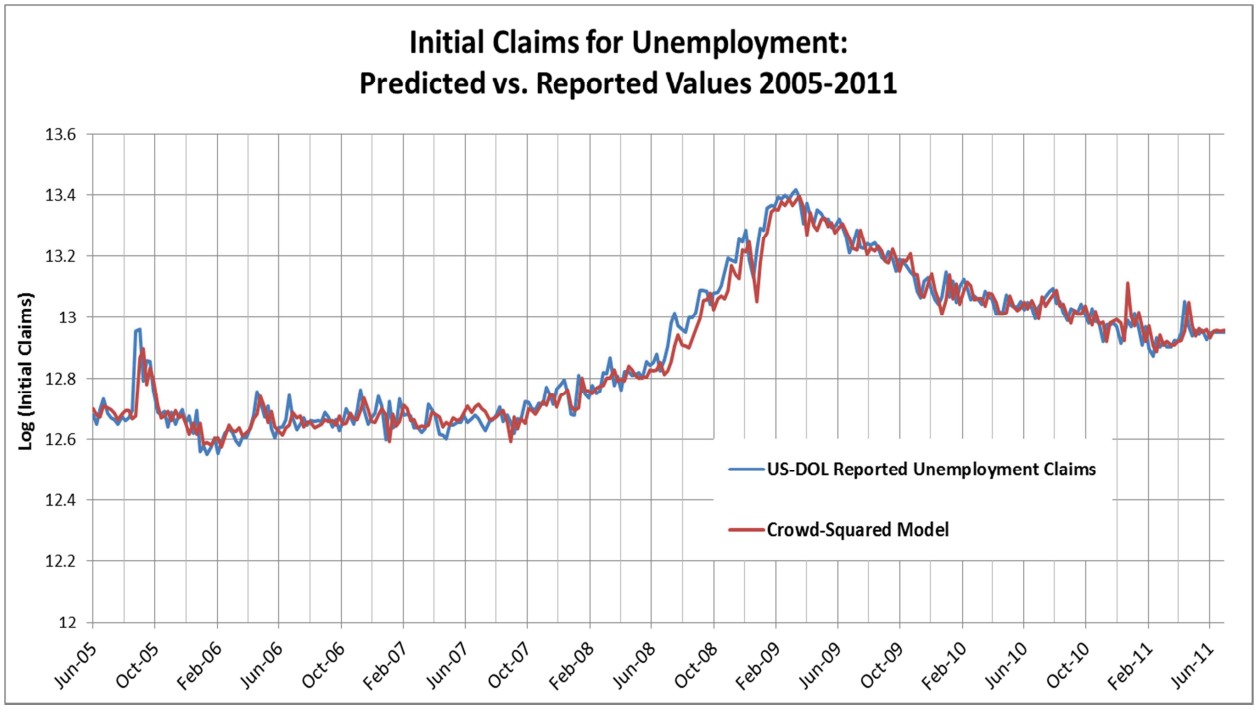


Figure 4. A comparison of the Association-based model predictions with the actual reported claims for unemployment published by the U.S. Department of Labor.

Discussion

Big data analytics allow researchers to perform real-time, large-scale monitoring of customer preferences and opinions; model and explain economic phenomena; and develop accurate predictions However, a critical aspect that hinders the utilization of large-scale crowd-based data for prediction is the lack of an effective method for selecting relevant data associated with the predicted item of interest.

This paper introduces crowd-squared, a new approach for using the crowd to identify relevant information in large-scale crowd-based data. Specifically, we used a word association game design to collect the associative thoughts of people to a focal phrase, a process that imitates the selection of search terms when using search engines. Thus we use one crowd to select the terms that a larger crowd will search for when seeking information about the phenomenon that we wish to predict.

We demonstrate this approach and show that even a straightforward implementation method can achieve improved prediction accuracy compared to categories hand-picked by expert researchers, and compared to high-power big-data technologies applied over large-scale and proprietary search log data.

We empirically tested our approach in three domains that were previously used for prediction generation (flu epidemics, housing market, and unemployment), and intentionally limited our analysis to the exact performance measures and data sets used in previous studies. Our results show that the crowd-squared keyword selection yielded a predictive dataset that outperformed data used previous published research. These results emphasize the importance of the keyword selection method in the prediction process, and demonstrate the robustness of utility of the crowd-square concept.

Managerial implications

Accurate measures of current events and predictions of future activities are one of the key challenges facing mangers and policy makers. The use of large-scale crowd-generated data has been shown to provide reliable estimates; However, their application to businesses has been hindered by the limitations of current term selection methods.

Our proposed approach may extend the potential use of search data for predictions, especially when the exact relevant keywords are unknown. Even when some prior knowledge exists, our proposed method can generate new related terms that can potentially improve predictive accuracy. Furthermore, due to its simplicity and low cost, forecasts can be updated periodically to support managerial decisions.

Methods implementing the crowd-squared concept can be used for both short-term and long-term decisions. For example, improved data selection can improve measurements of current demand trends, which in turn could assist in tasks such as shipment routing and planning of marketing activities in the short term. Improved data selection may also improve early detection of problems in current products or services. With respect to long-term decisions, more accurate predictions based on the crowd-square concept may facilitate more effective production planning, or reveal consumers' needs for product modifications or new products.

Overall, in the era of increasing volumes of big data, our approach allows for simple and low-cost filtering of relevant information that can be used in measurements and prediction of business activities.

Limitations and future research

While the use of search volume data has been shown to improve prediction models, it is important to note that people who perform online searches do not necessarily reflect a

representative sample of the population. For example, elderly people or people with low income tend to use the Internet less often, which could lead to inaccurate predictions in some domains. In addition, due to privacy constraints, Google makes search volume data available only when the number of searches of a specific term reaches a threshold that obstructs the possibility of using the aggregated data to identify the searchers. As a result, small-scale phenomena, or events that occur in areas with a low population density, will not be published by these search tools.

In a similar manner, the use of crowd-squared keyword selection tools may also fail to generate a representative sample of the population and may be unsuitable for areas with low populations or areas with a low level of technology adoption. Nevertheless, since crowd demographic properties can be collected in the crowd-squared process, this process can enable better matching of search terms to the target group whose behavior one wishes to predict. For example, a crowd of women between the ages of 20 and 25 may be used as the sample for keyword selection for sales predictions of a product that is commonly purchased by women of that age group. In future research we plan to analyze these types of demographic splits as a possible enhancement of crowd-squared based methods.

Finally, our analysis focused on search trend data whose simple structure makes it less prone to confounding factors such as the specific textual data processing method selected, and therefore enables accurate comparisons of predictive performance. One possible extension of this work could be analysis of predictive performance using social media data.

References

Archak, N., Ghose, A. and Ipeirotis, P. 2011. Deriving the pricing power of product features by mining consumer reviews. *Management Science*, 57(8), pp. 1485-509.

Bayus, B.L. 2013. Crowdsourcing new product ideas over time: an analysis of the Dell Ideastorm community. *Management Science* 59(1), pp 226-44.

Boudreau, K.J. and Lakhani, K.R. 2013. Using the crowd as an innovation partner. *Harvard Business Review* 91(4), pp 60-9.

Brabham, D.C. 2008. Crowdsourcing as a model for problem solving an introduction and cases. *Convergence: The International Journal of Research into New Media Technologies* 14(1), pp 75-90.

Chevalier, J. and. Mayzlin, D. 2006. The effect of word of mouth on sales: online book reviews. *Journal of Marketing Research* 43(3), pp. 345-54.

Choi, H. and Varian, H. 2012. Predicting the present with google trends. *Economic Record* 88(s1), pp. 2-9.

Chintagunta, P.K., Gopinath, S. and Venkataraman, S. 2011. The effect of online user reviews on movie box office performance: accounting for sequential rollout and aggregation across local markets. *Marketing Science* 29(5), 944-57.

D'Amuri, F. and Marcucci, J. 2012. The predictive power of Google searches in forecasting unemployment. *Bank of Italy Temi di Discussione (Working Paper) No* (891).

Dellarocas, C., Awad, N. and Zhang, X. 2007. Exploring the value of online product reviews in forecasting sales: the case of motion pictures. *Journal of Interactive Marketing*, 21(4), pp. 23–45.

Dewan, S. and Ramaprasad, J. 2012. Music blogging, online sampling, and the long tail. *Information Systems Research*, 23(3), pp. 1056–67.

Dhar, V. and Chang, E. 2009. Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing* 23 (4) pp. 300-307.

Duan, W., Gu, B. and Whinston, A.B.. 2008. The dynamics of online word-of-mouth and product sales—an empirical investigation of the movie industry. *Journal of Retailing*, 84(2), 233–42.

Du, R.Y. and Kamakura, W.A. 2012. Quantitative trendspotting. *Journal of Marketing Research* 49(4), pp. 514-36.

Geva, T., Oestreicher-Singer, G., Efron, N. and Shimshoni, Y. (2013), Do customers speak their minds? using forums and search for predicting sales. In *Proceedings of the 2013 International Conference on Information Systems*.

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* 457(7232), pp. 1012-14.

Godes, D. and Mayzlin, D. 2004. Using online conversations to study word-of-mouth communication. Marketing Science, 23(4), pp. 545-60.

Goel, S., Hofman, J.M., Lahaie, S., Pennock, D.M. and Watts, D.J. 2010. Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences* 107(41), pp. 17486-90.

Howe, J. 2006. The rise of crowdsourcing. *Wired Magazine* 14(6), pp. 1-4.

Lakhani, K.R., Jeppesen, L.B., Lohse, P.A. and Panetta, J. A. 2007. *The value of openness in scientific problem solving*. Boston, MA: Harvard Business School (working paper).

Liu, H., Motoda H., 1998. Feature selection for knowledge discovery and data mining. Springer.

Liu, Y. 2006. Word of mouth for movies: its dynamics and impact on box office revenue. *Journal of Marketing* 70(3), pp. 74–89.

McAfee, A. and Brynjolfsson, E. 2012. "Big data: the management revolution," *Harvard business review* October 2012, pp 2-9.

Moe, W. W. and Fader, P. S. 2004. Dynamic conversion behavior at e-commerce sites. Management Science, 50(3), 326-335.

Nelson, D.L., McEvoy, C.L. and Dennis, S. 2000. What is free association and what does it measure? *Memory & Cognition* 28(6), pp. 887-99.

Nelson, D.L., McEvoy, C.L. and Schreiber, T.A. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers* 36(3), pp. 402-407.

Nelson, D.L., McKinney, V.M., Gee, N.R. and Janczura, G.A. 1998. Interpreting the influence of implicitly activated memories on recall and recognition. *Psychological Review* 105(2), p. 299.

Netzer, O., Feldman, R., Goldenberg, J. and Fresko, M. 2012. Mine your own business: market-structure surveillance through text mining. *Marketing Science*, 31(3), pp. 521-43.

Seebach, C., Pahlke, I., and Beck, R. 2011. Tracking the digital footprints of customers: how firms can improve their sensing abilities to achieve business agility. In: *ECIS 2011 Proceedings*.

Snow, R., O'Connor, B., Jurafsky, D. and Ng, A.Y. 2008. Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In:

Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics 2008.

Sparrow, B., Liu, J. and Wegner, D.M. 2011. Google effects on memory: Cognitive consequences of having information at our fingertips. *Science*, 333 (6043), pp 776-8.

Rui, H., Liu, Y. and Whinston, A. 2013. Whose and what chatter matters? the effect of tweets on movie sales. *Decision Support Systems* 55(4), pp. 863-70.

Von Ahn, L. 2006. Games with a purpose. *Computer* 39(6), pp 92-4.

Vosen, S. and Schmidt, T. 2011. Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting* 30(6), pp. 565-78.

Wu, L. and Brynjolfsson, E. 2009. The future of prediction: how Google searches foreshadow housing prices and quantities. In: *Proceedings of the 30th International Conference on Information Systems*, Phoenix, Arizona.

Yan, T., Kumar, V. and Ganesan, D. 2010. Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In: *Proceedings of ACM2010 - The 8th international conference on Mobile systems, applications, and services*.

Appendix A – List of Aggregated Associated Terms

Table 1.

Top 10 Associated Terms by Cue Term

Influenza		**Housing Sales**		**Unemployment**	
Term	*Association strength*[*]	*Term*	*Association strength*	*Term*	*Association strength*
sick	53%	mortgage	50%	poor	20%
fever	47%	expensive	18%	money	20%
cold	19%	realtor	18%	jobless	16%
cough	18%	location	16%	depression	16%
contagious	15%	money	14%	broke	12%
germs	11%	loan	12%	homeless	12%
shot	10%	agent	8%	bills	10%
vaccine	10%	interest rate	8%	no money	10%
influenza	9%	real estate	8%	sad	10%
virus	9%	bank	8%	economy	8%

* *Association strength* is the percentage of participants providing this word.