Using Crowd-Based Data Selection to Improve the Predictive Power of Search Trend 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

Large-scale data generated by crowds provide a myriad of opportunities for monitoring and modeling people's intentions, preferences, and opinions. A crucial step in analyzing such "Big Data" is identifying the relevant data items that should be provided as input to the modeling process. Interestingly, this important step has received limited attention in previous research. This paper proposes a novel crowd-based approach to this data selection problem: leveraging crowds to amplify the predictive capacity of search trend data (Google Trends). We developed an online word association task that taps into people's "thought-collection" process when thinking about a focal term. We empirically tested this method in two domains that have been used as test-beds for prediction. The method yields predictions that are equivalent or superior to those obtained in previous studies (using alternative data selection methods) and to predictions obtained using various benchmark data selection methods.

Keywords: Prediction, Search Trend, Big Data
Introduction

Digital data provide opportunities to monitor customers’ intentions, preferences and opinions in real time and on a massive scale. As a result, it has become possible to model and explain economic phenomena that previously defied prediction. Yet the abundance of data also creates significant challenges for data collection and processing.

One of the most important challenges that has emerged is determining which data items should be selected for modeling a phenomenon of interest. Given the vast quantity of available data (for example, all possible search engine queries, literally numbering in the billions), it is essential to select the specific data items that are relevant for a given modeling or prediction task. Clearly, excluding relevant data will render a modeling process “blind” to important information; in contrast, including irrelevant data can confound the modeling process and can result in undesired outcomes such as overfitting, in addition to creating additional processing and data collection costs.

In this paper, we offer a novel crowd-based approach to this data selection problem: leveraging crowds to amplify the predictive capacity of search volume data (e.g. Google Trends). We present a simple, inexpensive implementation that demonstrates the predictive capacity of our approach in comparison to data selection methods used in previous studies. Specifically, our demonstration uses Google Trends1 and an online task designated for a crowdsourcing environment.

Our method is based on prompting a large number of individuals, via an online task, to produce word associations relating to a focal term (reflecting the phenomenon for which predictions are being made). This approach taps into people’s lexical knowledge (Nelson et al. 2004) and provides a relative index of the accessibility of related words in individuals’ memory. A key element in of the association game is the fact that it provides a power law distribution of term associations (Steyvers and Tenenbaum 2005)—most associations relate to terms that are proximal to the cue term, yet a “long tail” of associations connect to more distant terms. Thus, this technique is expected to provide ample coverage of relevant data items.

As illustrated in Figure 1, data selection (step 2) is a critical stage in the process of producing predictions on the basis of large-scale online crowd-generated data. This stage bridges between the generation of data by crowds and extraction of actual inputs for modeling. In particular, it is the first stage in which the modeler actively makes decisions.

Figure 1 also highlights the distinction between data selection (step 2) and the well-known feature selection problem (Liu and Motoda 1998) (step 5).2 Specifically, the data selection task addresses the question of which data, out of all possible data that potentially could be collected, should actually be collected by the researcher. In contrast, the feature selection process (if applied) takes place at a later stage, typically after the selected data have undergone further processing. In particular, it deals with the selection of a subset of informative variables out of a larger set, generated from the collected data. Therefore, the feature selection process is totally dependent on the data selected in step 2 and is effectively blind to data that were not selected in that step.3

---

1 Google Trends (http://www.google.com/trends/) is a publicly available product that aggregates billions of search queries and provides information about the relative volume of different search terms. This information is currently available under some restrictions (for example, only several hundred search trends can be obtained on a daily basis).

2 We note that regularization (Bishop 1995), which “downplays” features rather than completely eliminating them, is an alternative to feature selection. Yet, like feature selection, this process is conducted at a later stage than the data selection stage and is effectively blind to data that were not selected during the data selection phase.

3 In the methodology section we discuss how feature selection procedures could be applied as a subsequent and complementary step to our data selection methodology.
Interestingly, the important step of data selection has received only limited attention in previous literature. This step is commonly carried out using three main approaches: (1) a comprehensive scan of large scale data, (2) intuition and prior knowledge, and (3) algorithmic classification methods. While these methods have been shown to provide good results in various domains, they also suffer from various downsides including: spurious correlation (as described by Lazer et al. 2014), being domain-specific, having strong dependence on the researcher’s prior knowledge, high computational costs, and reliance by the researchers on proprietary data.

After explaining the novel methodology of our data selection method, we use it to identify relevant terms, collect the corresponding search trend data, and generate predictions in two different domains: influenza epidemics and unemployment claims. We then compare our results with a challenging and well-known benchmark model in each domain as well as with alternative benchmark methods. We find the use of the crowd-based method to be highly effective. Our results suggest that the integration of crowd-selected search terms with aggregated data from search engines performs as well as these benchmarks or even outperforms them—and does so at a very low modeling cost. Additional advantages of our methodology are robustness across long periods of time, improved understandability, and finer-grained analysis capabilities compared to several benchmark methods.

To the best of our knowledge, this paper is the first to propose a structured approach for data selection in large-scale crowd-generated data environments. By using the crowd to address the data selection process we are able to capture a set of terms that are both relevant and predictive. Another benefit of our method is the transparency of the selection process and ease of replication. These are important traits that, according to Lazer et al. (2014), do not typically receive sufficient attention in studies analyzing large-scale data.

**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 carry out 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.

Early work in this domain includes studies using social media data to explain and predict various economic outcomes (e.g. Godes and Mayzlin 2004). The success of these studies resulted in significant momentum to this research field, which has since generated a large number of scientific studies.

Research using social media data is not the only stream of research that has used crowd-generated data to change the way we predict events. In particular, search engine logs or search trends, aggregating large
quantities of crowd-generated search queries, have begun to receive significant attention for their utility in detecting and predicting a variety of economic outcomes. Ginsberg et al. (2009) were the first to show that search volume data can be leveraged to generate useful predictions of influenza epidemic outbreaks. This breakthrough paper opened the door to a stream of studies that used such online activity data to generate predictions in a wide range of domains, including movie box office sales and music billboard rankings (Goel et al. 2010), 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 Schmidt 2011).

However, even with the availability of big data technologies in addition to 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. Data selection, and particularly selection of relevant data terms, is a challenging task. Online data items might relate to a specific phenomenon using several terms (e.g., influenza may also be referred to as "flu" or "cold"). In straightforward cases, the terms 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. It is also possible that the relevance of a keyword to a phenomenon of interest may be even less direct. For instance, Wu and Brynjolfsson (2009) showed that home purchases are indicative of future home appliance sales.

**Data Selection Approaches**

As noted above, data selection is typically performed using three main approaches: (1) a comprehensive scan of large scale data, (2) intuition and prior knowledge, and (3) algorithmic classification methods.

The first approach for data selection uses a comprehensive scan of all available data to select the terms that are most strongly correlated with the focal phenomenon. Ginsberg et al. (2009), for example, used Google’s internal data 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. As pointed out by Lazer et al. (2014), one of the main problems of this approach is overfitting due to a spurious correlation. When correlating enormous number of search terms with a small number of data points there is a high likelihood of finding search terms that are correlated with past occurrences of the focal phenomenon but unrelated to the phenomenon itself and therefor would have no predictive power.

Additionally, as in the case of Ginsberg et al. (2009), this data selection approach commonly resorts to the use of proprietary data. Thus, it is impossible to reproduce this methodology using the limited data that are publicly available; Google, for example, imposes a strict limitation on the number of terms that external users can extract from Google Trends (several hundred per day). Furthermore, the analysis performed by Ginsberg et al. (2009) required expertise and computational power to create the correlation matrix for the phenomenon of interest.

Using the second approach, the researcher applies human intuition and prior knowledge to identify online data pertaining to a certain item (e.g., flu). This task commonly involves choosing terms that are likely to be associated with that item (e.g., “flu”, “influenza”). This approach has been employed in various studies, using straightforward keywords. For instance, Dhar and Chang (2009) used music album titles and Rui et al. (2013) used movie name mentions on Twitter to identify word of mouth relating to the mentioned products. Other studies such as D’Amuri and Marcucci (2012) used similarly straightforward terms in conjunction with search trend data. Naturally, the modeling outcomes using this approach are domain-specific, and therefore it is difficult to use formal methodology to represent researchers’ individual knowledge.4

---

4 Another variant of reliance on "previous researcher knowledge" for data selection appears in various studies that choose data from webpages in which an item of interest is clearly identified by the researchers. See, for instance, Chintagunta et al. (2011), Dellarocas et al. (2007), Dewan and Ramaprasad (2012), Duan et al. (2008), and Liu (2006), who used user reviews from dedicated webpages for specific movies or songs. While focusing exclusively on clearly identifiable data is suitable for various research goals, this approach is blind to data from many other
The third approach uses automated methods to classify data into predefined categories. For instance, Google Trends offers an internal "black box" category classifier, which has been used in various studies such as Choi and Varian (2012), Wu and Brynjolfsson (2009), and Vosen and Schmidt (2011). These classifiers commonly focus on detecting data items that pertain to a pre-determined category (e.g., search queries that relate to the automotive industry). 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 (as in the above mentioned example of Wu and Brynjolfsson, 2009, who found that search trends related to new home purchases were indicative of future home appliance sales). Additionally, these classifiers are not transparent to the typical end user (i.e., a user who is not affiliated with the classifier developers), who might consequently find it difficult to gauge their accuracy or coverage. In fact, to some extent, this third approach may be a variant of the second approach, as it is possible that a classifier’s rules are determined (or that examples are provided for a supervised learning-based method) on the basis of the developer’s prior intuition.

**Crowdsourcing**

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 the evaluation of our methodology, is that search query texts are commonly brief and focused. As a result, the use of search trends enables us to develop a direct application of our proposed method and to avoid the use of various intervening procedures that are required to extract terms from complex texts.

The fundamental idea underlying prediction based on search trend 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 choose 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.

We present a new technique that uses crowdsourcing to generate the relevant keywords that could reflect consumers’ intentions. 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, crowdsourcing has become increasingly prevalent in many fields, and is used in a variety of 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 crowd-sourcing is finding methods of engaging 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 a single task from multiple non-expert individuals can generate results at the same level as those created by experts.

We note that the term-selection task we propose shares certain features with query expansion methods (such as relevance feedback) reported in information retrieval literature (see, for instance, Manning et al. 2008), which, too, seek to identify sets of query terms that are suitable for a given purpose. However, there are several important aspects in which our approach differs substantially from query expansion methods. First, the goal of our method is to select terms that are indicative of relevant, crowd-generated data items, for predicting a real-world phenomenon. Query expansion methods, in contrast, aim to provide a set of terms that can subsequently retrieve a set of documents (out of a larger collection of webpages that may contain relevant information (e.g., it is possible that a new posting on an actor's web page may provide additional information on the box office performance of her recent film).
documents) that contain particular information that a person seeks. For example, in the context of search engine data, our crowd-based method could be used to find a set of search terms whose corresponding volume, generated by millions of individuals, is predictive of real-world flu outbreaks. Query expansion methods, in contrast, would be used to identify terms that can retrieve a set of documents that contain information about the flu.

Given their different underlying goals, the two approaches are clearly expected to yield different types of query terms. For example, in identifying terms corresponding to internet searches predictive of real-world flu outbreaks, our approach would also seek to include terms that capture data generated by people who are unaware that they are searching for information about the flu (e.g., people searching for flu symptoms such as fever or cough, without knowing they have the flu). In query expansion methods, in contrast, user awareness of the search goal is commonly required.

Finally, to employ our method, there is no need to obtain the actual data collection (e.g., a scan of the internet), or individual data queries. In contrast, query expansion methods such as relevance feedback commonly require access to such data—they are based on iterative procedures that evaluate the user's queries and subsequent document selections (or user document clickstream).

In the following sections we elaborate on and demonstrate a new crowd-based approach that uses an online word association game that captures people’s ideas regarding terms that relate to focal phrases. We subsequently show that the resulting terms can be used to identify relevant data for predictions in two different domains.

**Methodology**

As detailed above, data selection based on potentially relevant terms (or keywords) is a common practice. A key characteristic of such term-based data selection is an inherent tradeoff between keyword coverage and accuracy (Geva et al. 2013). For instance, the use of a large set of keywords may provide ample coverage of all relevant search trend data (i.e., it might yield a data set comprising all search trends that are relevant for modeling a certain phenomenon). However, using such a large set of keywords might come at a cost of including some irrelevant terms, which, in turn, may confound the prediction model or result in overfitting the data.

The data selection process we propose (step 2 in Figure 1) aims primarily to provide high coverage of relevant data. To efficiently balance this coverage with accuracy, we supplement the data selection stage with a secondary step: a feature selection process (step 5 in Figure 1).

**Crowd-Sourced Data Selection**

For the primary data selection step of our methodology we introduce a technique to use crowd-sourced 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 or terms.

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). Individuals rely on associations in everyday activities as a means of “collecting thoughts” (Nelson et al. 2000).

Word association provides an index of the probability that words are related to the cue term. This information has been 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 form of external or transactive memory (Sparrow 2011), word association can be used to determine effective search queries. With its

---

5 Our definitions of the terms "accuracy" and "coverage" are similar to those used in Geva et al. (2013): Let $K$ denote a set of keywords, and $P$ denote a phenomenon of interest. “Accuracy” denotes the ratio between the number of data items (search queries) that specify (any word in) $K$ and that actually relate to $P$, and the total number of data items (search queries) specifying any word in $K$. “Coverage” denotes the ratio between the number of data items (search queries) using any word in $K$, and the hypothetical, full number of data items (search queries) referring to $P$ (using any keyword).
consistent representations of the associated terms, these terms may reflect broader search patterns and therefore assist in measuring current events and predicting future activities.

Specifically, a key benefit of the word association technique is that it provides a power law distribution of term associations (Steyvers and Tenenbaum 2005). While most associations relate to terms that are proximal to the cue, a “long tail” of associations connects to more distant terms. Thus, this technique allows us to capture many potentially relevant terms that are not necessarily highly correlated with one another, thereby providing wide coverage of data items that may pertain to the phenomenon of interest.

**Word association game design**

We designed and developed an online word association website specifically 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, as elaborated below (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.

We used the Amazon Mechanical Turk platform\(^6\), an online marketplace for tasks that require human intelligence (or tasks that are easily achieved by a human but require large computational 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 task assignments to be randomly assigned multiple Turkers and provides control over task completion.

In total, 1,100 Turkers participated in our experiments (550 for each domain).\(^7\) Each participant (Turker) was given a single cue phrase and was asked to provide five terms or phrases that came to mind when seeing this phrase. Each Turker was paid 5-8 cents ($0.05-$0.08) for completing the game. The average duration of a game was 53 seconds (including completion of 4 demographic items).

The resulting set of different associated phrases for each domain was very large. Nevertheless, it is important to account for the fact that the use of any single phrase may not represent a common form of thinking, reflective of multiple individuals’ search patterns, but rather might be indicative of only one individual’s unique thinking. As shown by Snow et al. (2008), an aggregation of results from multiple individuals can generate results of high quality. We therefore restricted the analysis to include only terms reported by at least 1% of the users.\(^8\) For each term, we then collected its search query volume over time and subsequently used these data in the modeling process.\(^9\)

---

\(^6\) Amazon Mechanical Turk (AMT) was used as a convenient mean to reach a large number of people and is not an essential part of our approach. Other platforms could also be used in a similar way to collect the keywords from the crowd.

\(^7\) We eliminated from the analysis 15 Turkers, who did not provide 5 associated terms or did not answer all demographic questions.

\(^8\) Google Trends has a policy intended to preserve privacy and therefore does not return results for certain keywords with relatively low search volume. It also rounds numbers to integer value; thus, it effectively rounds down low search query volume to zero. We therefore exclude from our set of keywords those terms for which Google Trends did not return results or that included zero in more than 90% of the instances in the training set. Due to space limitations the full list of “valid” terms is not included in this paper. It is available upon request from the authors.

\(^9\) As detailed in Choi and Varian (2012), Google Trends data are computed by a sampling method and therefore may contain some noise. To reduce the noise we use a similar procedure as in Preis et al. (2013) of averaging the value of multiple draws from Google Trends.
Feature Selection

As detailed above, the crowd-sourced data selection method is designed to use the power-law distribution characterizing association games to select data with high coverage. Nevertheless, this method of data selection might introduce into the resultant data set some search trends with low accuracy (i.e., with low relevance to the phenomenon of interest). Therefore, as noted above, we implemented a feature selection procedure to choose the most useful variables and thereby to balance the inherent tradeoff between coverage and accuracy within our data.

Feature selection procedures have been widely reported in the literature and are generally known for their ability to improve predictive accuracy, reduce overfitting and decrease model complexity (Liu & Motoda, 1998). Our implementation of feature selection involves a standard sequential forward selection procedure with a nested holdout sample (see, for instance, Provost and Fawcett 2013). Specifically, we train prediction models over the first 2/3 of a training set and repeatedly add the next most useful feature (search volume for a given keyword), to the prediction model based on its contribution to reducing the Mean Absolute Error (MAE) criteria over a nested holdout sample that consists of the last 1/3 of the training data instances. We repeat this procedure for up to 30 features and then select the set of features that optimize the relevant performance criteria over the nested holdout sample. After selecting the number of relevant features, the model is then re-trained over the entire training set data and performance is evaluated over an external, independent, validation set, which consists of data items from a subsequent time period.  

We note that our main goal is to evaluate the efficacy of our primary stage of crowdsourcing the data selection process; we use the feature selection only as a complementary step to eliminate data items with low accuracy. Therefore, while literature offers a host of advanced feature selection methods, we intentionally apply a straightforward method. Thus, our implementation of the crowd-sourced data selection methodology in conjunction with a straightforward feature selection method is designated to provide a rather conservative assessment of the performance of the overall data selection methodology. Clearly, if this method is found to be successful, then applying more advanced feature selection methods could obtain even better results.

---

As detailed in the evaluation section – we constrain ourselves and use the same time periods for training the model as in Ginsberg et al. (2009) and Choi and Varian (2012). The nested holdout sample is a subset of these training periods. Overall predictive capacity in the evaluation section is measured using the same, external validation set reported in these studies.
Evaluation

To validate the effectiveness of the crowd-sourced data selection approach, we applied our proposed methodology for prediction tasks in two different domains. Specifically, we replicated tasks reported in two well-known related studies: Ginsberg et al. (2009) in the detection of influenza outbreaks, and Choi and Varian (2012) in predictions of unemployment levels.

To ensure 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. Furthermore, in the domain of unemployment levels, we also compared our results with those of a baseline model evaluated by Choi and Varian (2012).

We further suggest an additional, external, benchmark for our methodology that relies on an alternative, simple-to-use data selection method for search trend data. Specifically, we use Google Correlate (see https://www.google.com/trends/correlate), an online tool that retrieves up to 100 terms according to the correlation of their search volume with an input time series. Google Correlate also ranks the importance of search terms, displaying the correlation values with the input time series. As the ease of use associated with this alternative method is comparable to that of our methodology, we believe it may serve as an additional valid benchmark for our method.

Influenza epidemics

The first data set that we used to validate our methodology is flu outbreak data from the U.S. Center for Disease Control (CDC). Ginsberg et al. (2009) used this type of data to construct an early detection system for influenza epidemics (named "Google Flu Trends"). Specifically, the dependent variable in their study was the weekly Influenza-Like Illness (ILI) factor reported by the CDC. To select the search term data 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. Ginsberg et al. 2009 reported that 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 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 January 2004 and the week commencing on March 11, 2007. 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. (2009).

Using the word association setting described above, we asked 535 Turkers (40% female, average age 29.1) to play the online game, where the task description was “Please write 5 terms that come to mind when seeing the word ‘Flu’.”

For each phrase reported by at least 1% of the Turkers, we collected the weekly search index from Google Trends (overall 72 valid terms were identified). 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 U.S. to match the predicted variable: flu outbreak in the U.S. For the sake of fair comparison, similarly to Ginsberg et al. 2009, we also used a linear prediction model.

11 After obtaining the search trend data recommended by Google Correlate – in order to determine the “optimal” number of Google Correlate benchmark-based features to be included in the model, we sequentially added the top 1-30 ranked Google Correlate-based recommended search trends, and evaluated their predictive performance using a similar linear prediction model over the same “nested validation” data reported above. We then selected the best-performing feature combination. We note that while the data items retrieved by Google Correlate are already ranked according to their correlation with the dependent variable—for robustness, we also tested whether applying the same feature selection procedure we used for the crowd-sourced data selection process would improve the Google Correlate-data-based prediction results. However, applying this feature selection procedure did not improve the Google Correlate benchmark performance, and therefore these results are not reported.

12 We excluded data from 2003 since Google Trends provides data only from 2004.
Specifically, we used the following prediction model:

\[ ILI(t) = \alpha + \sum_i \beta_i \text{AssociatedTerm}_i(t) + \epsilon_i \]  

(1)

Where \( ILI(t) \) is the ILI factor at time \( t \) reported by the CDC; \( \text{AssociatedTerm}_i(t) \) is the search trend value at time \( t \) for the association-based term \( i \) in the aggregated results of the word association game for influenza.

Table 1 summarizes the results for the different data selection methods. As shown in this table, our prediction results achieved a very similar level of out-of-sample correlation (0.966) in predicting ILI (compared to 0.97 in Ginsberg et al., 2009). Given that the two data sets yielded comparable 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. (2009) used 50 million different search terms and 450 million different models to generate the final model, which 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 72 valid terms suggested by 535 online users; following our suggested secondary feature selection procedure, our model ultimately used only 22 terms.

Table 1 – Correlation results with CDC-reported ILI for prediction models using different data selection methods

<table>
<thead>
<tr>
<th>Validation Period</th>
<th>Crowd-Sourced Data Selection</th>
<th>Google Flu Trends</th>
<th>Google Correlate</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2007 – May 2008 (similar to Ginsberg et al.)</td>
<td>0.966</td>
<td>0.97</td>
<td>0.965</td>
</tr>
<tr>
<td>April 2012–March 2013</td>
<td>0.97</td>
<td>0.951</td>
<td>0.954</td>
</tr>
</tbody>
</table>

While our data selection method achieved comparable results to those of Ginsberg et al. (2009) over the same validation period used in that study (as presented in Table 1 and Figure 3), a recent study by Lazer et al. (2014) suggests that Google’s Flu Trends prediction system (which is based on Ginsberg et al. 2009) has been modified over recent years to account for new data, yet its performance deteriorated, especially during the major flu outbreak of 2012-2013. Therefore, it is interesting to evaluate the robustness of predictions based on our data selection methodology during this more recent period. This setting is especially challenging due to the fact that our model was trained on much earlier data, from 2004-2007.

Specifically, we compared our results, with the results provided by the Google Flu Trends website.\(^{13}\) Here, our results show a noticeable improvement in correlation levels compared to the Google Flu Trends system for the period April 2012 and March 2013. Specifically, our method obtained a correlation of 0.97, whereas the Google Flu Trends prediction obtained a correlation of 0.951. It is also interesting to note that, unlike the results of the Google Flu Trends system, prediction results based on our data selection method remained stable over time and did not deteriorate.

\(^{13}\) http://www.google.org/flutrends/us/data.txt.
Furthermore, 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 than did the predictions of the Google Flu Trends model.

Figure 3. A comparison of the crowd-sourced data selection model with actual reported ILI and Ginsberg et al. (2009)/Google Flu Trends, over two separate periods: 2007-2008 (Top) and 2012-2013 (Bottom).

Furthermore, 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 than did the predictions of the Google Flu Trends model.

We normalized the data to allow for a consistent comparison in Figure 3, since the data available on Google Flu Trends are on a different scale. All values were normalized by dividing the value in a specific week by the respective data series average.
Last, we compared our results to the results obtained by another benchmark data selection method: choosing search trends based on terms provided by Google Correlate. Overall, we found that, for the validation period reported in Ginsberg et al. (2009), predictions based on our data selection methodology slightly outperformed those based on Google Correlate data selection. Yet, for the years 2012–2013, our methodology considerably outperformed predictions using Google Correlate-based data selection. Specifically, Google Correlate obtained a correlation of 0.965 for the validation period reported in Ginsberg et al. (2009) and a correlation of 0.954 for the year 2012-2013, whereas the corresponding correlation values obtained with the crowd-sourced method were 0.966 and 0.97, respectively.

**Flu prediction sensitivity analysis: number of keywords**

The results reported above for the crowd-sourced data selection method used a final set of search trends that yielded the best performance over a nested validation set (last third of the training set)—as determined by the feature selection procedure detailed in the methodology section. For robustness, we provide in Figure 4 a detailed analysis of predictive capacity when constraining this feature selection process to select a fixed number of search trends—from a single feature and up to 30 features. This figure also presents the results of imposing a similar constraint on the number of search trends obtained from Google Correlate. This figure shows that when 6 or more search trends are used, the predictive capacity of the crowd-sourced data selection method for the 2007–2008 validation period is comparable to that of Google Flu Trends and to that of the Google Correlate-based data selection method. Notably, the crowd-sourced method considerably outperforms these methods in the more challenging 2012-2013 flu outbreak season.

![Correlation results while using a fixed number of features](image)

**Figure 4 – Correlation results while using a fixed number of features**

2007-2008 validation (Left) and 2012-2013 validation (Right)

15 We provided to Google Correlate tool as an input the dependent variable (ILI CDC reports) data, during the same training set period.
Initial claims for unemployment benefits

The second set of data that we used to evaluate the capacity of the crowd-sourced data selection methods involves the task of early assessment of the volume of initial claims for unemployment benefits in the U.S. 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. Choi and Varian (2012) carried out a study that used search trend data to produce early assessments of initial claims for unemployment. They developed a 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 expanding window prediction (that is, using the data up until week \((t-1)\) to train the model and measure its performance over week \((t)\)), using training and validation data from January 2004 to July 2011.

Their model was able to generate accurate predictions of economic turning points, and their overall result, measured by Mean Absolute Error (MAE), was 3.68%. However, Choi and Varian also reported that a simple baseline model that did not incorporate search trend data generally outperformed the model they developed, yielding an MAE of 3.37% over the measured time period. This model is presented in equation (2):

\[
UIC(t) = \alpha + \delta_t UIC(t - 1),
\]

where \(UIC(t)\) is the logarithm of the seasonally-adjusted volume of initial claims for unemployment for week \(t\).

This result suggests that the search trend data, based on the predefined categories used by Choi and Varian, might have contained information that overlapped with the information contained in the data on the previous week’s claims, in addition to some noise that may have reduced out-of-sample predictive accuracy.

To generate our data set, we asked 545 Turkers (58.5% female, average age 33) to perform the word association task described above. In this case, the task description was “Please write 5 terms that come to mind when seeing the phrase ‘Unemployment’.” For each phrase reported by at least 1% of the Turkers, we collected the weekly search index from Google Trends (overall, 91 valid terms were identified).

We replicated the exact processing and modeling steps reported in Choi and Varian (2012), including the same process of adjusting for seasonality as well as using an expanding window, one-step-ahead prediction methodology.

Specifically, we used the list of search trends derived from our data selection procedure and reran a simple linear regression model as detailed in equation (3) for each one-step-ahead prediction:

\[
UIC(t) = \alpha + \delta_t UIC(t - 1) + \sum \beta_i AssociatedTerm_i(t),
\]

where \(UIC(t)\) is the logarithm of the seasonally-adjusted volume of initial claims for unemployment for week \(t\), and \(AssociatedTerm(t)\) is the search trend value for the association-based term \(i\) at week \(t\).

We applied this model to the same period similar used in Choi and Varian (2012) (see Figure 5 for a comparison of the prediction model and actual unemployment claims data). Our prediction model obtained an out-of-sample MAE value of 3.23%. This value is better than the MAE value for the competent baseline model (3.37%) and is also superior to the MAE value reported by Choi and Varian (2012) (3.68%). This demonstrates that the crowd-sourced data selection method can outperform a

---


17 Choi and Varian (2012) report that it is challenging to obtain better predictions than with an AR1, since the data are likely to represent a random walk (with a drift) behavior.

18 To maintain consistency with Choi and Varian (2012), we also seasonally adjusted the associated terms using the STL function in R. Additionally, since Choi and Varian used an expanding window methodology – our feature selection procedure was repeated for each expanding window iteration.
competent baseline model even when other methods such as using Google’s automated classifier (combined with experts, hand-picking the classifier categories) fail.

Last, for robustness, we compared our results to the results of an additional, external benchmark—choosing search trends based on terms provided by the Google Correlate tool. We found that predictions based on our data selection methodology (obtaining an MAE of 3.23%) also outperformed Google Correlate-based search trends selection (which obtained an MAE of 3.47%).

Unemployment prediction sensitivity analysis: number of keywords

For robustness, we provide in Figure 6 a detailed analysis of constraining this feature selection process to select a fixed number of search trends from a single feature up to 30 features. This figure also presents the outcome of imposing a similar constraint on the number of search trends based on data selected by Google Correlate. This figure shows that when 4 or more search trends are used, the crowd-sourced data selection method consistently obtains better results than do all 3 benchmarks.
Conclusions

Accurate monitoring of current events and predictions of future activities are key challenges faced by managers as well as by researchers. The use of search trend data has been shown to provide fairly accurate estimates. However, a critical aspect that hinders the use of this type of data for prediction is the lack of an effective method for selecting relevant data associated with the predicted phenomenon of interest.

To the best of our knowledge, this paper is the first to propose a structured approach for data selection in search trend data environments. Specifically, we used a word association game designed to collect individuals’ associative thoughts in relation to a focal phrase, with the goal of obtaining high coverage of relevant data items. Thus, we used one crowd to generate a high-coverage list of terms representing the keywords that a larger crowd is expected to search for when seeking information about a phenomenon that we wish to predict.

We demonstrate an application of this approach and show that even a straightforward implementation method can achieve similar or better prediction accuracy compared with categories hand-picked by expert researchers, high-power big-data technologies applied over large-scale and proprietary search log data, as well as compared to additional benchmarks. We achieved this improved accuracy while intentionally constraining our predictive analysis to use the exact same algorithms, data sets and performance measures used in previous well-known studies. Our results emphasize the importance of the data selection method in the prediction process, and demonstrate the utility of the crowd-sourced data selection concept.

Furthermore, the successful performance of the crowd-sourced data selection method across different domains and in comparison to different benchmarks indicates its robustness. Accurate prediction is crucial in many business applications. Furthermore, even small improvements in prediction quality can have enormous economic and social benefits. Thus, we expect that this method will have managerial implications that go beyond the specific domains used in this study. In addition, our proposed approach may extend the potential use of search data for predictions, especially when the exact relevant keywords
are unknown. Even in cases in which some prior knowledge exists, our proposed method can generate new related terms that can potentially improve predictive accuracy. Due to the method’s simplicity and low cost, forecasts can also be updated periodically to support managerial decisions.

**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 constitute 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 a crowd for keyword selection tools may also fail to generate a representative sample of the population and may be unsuitable for areas with small populations or areas with a low level of technology adoption. Nevertheless, since crowd demographic properties can be collected in the crowd-sourced process, this process can enable search terms to be matched 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-based methods.

Finally, our analysis focused on search trend data, whose simple textual structure enabled us to avoid the use of various intervening procedures that are required to extract terms from complex texts. In doing so, we avoided introducing confounding factors and were able to carry out accurate comparisons of predictive performance. Nevertheless, one possible extension of this work could be analysis of predictive performance using various types of social media data.
References


