Soc. Netw. Anal. Min. (2015) 5:7
DOI 10.1007/s13278-014-0239-8

ORIGINAL ARTICLE

Health-related hypothesis generation using social media data

Jon Parker · Andrew Yates · Nazli Goharian · Ophir Frieder

Received: 31 December 2013 / Revised: 5 November 2014 / Accepted: 14 November 2014

Abstract Traditional public health surveillance, also known as syndromic surveillance, is expensive and burdensome because it relies on clinical reports authored by health professionals with considerable time and effort. Due to its preventative cost, syndromic surveillance is typically only performed for high risk concerns like influenza. Therefore, a health surveillance system that works for numerous health concerns simultaneously would be of great practical use. We present a framework that processes a stream of time-stamped social media messages. The framework produces "interest curves" that permit the generation of hypotheses regarding which health-related conditions/topics may be increasing in prevalence. We do not claim to detect an actual outbreak of a health-related condition because this framework only has access to social media messages and not a harder data source like patient records. This approach differs from other prior approaches because it is not customized to detect one particular illness (e.g., influenza) as is commonly done. The inner workings of the framework can be interpreted as a transformation that converts a signal deeply embedded in the "stream of raw tweets" domain to a signal in the "health related topics" domain. This framework's capability is demonstrated by examining multiple interest curves related to seasonal influenza and allergies.

Keywords Twitter · Health surveillance · Trend detection · Item-set mining · Wikipedia

J. Parker (✉) · A. Yates · N. Goharian · O. Frieder
Information Retrieval Laboratory, Department of Computer Science, Georgetown University, Washington, DC, USA
e-mail: jon@ir.cs.georgetown.edu; jparker5@jhmi.edu

A. Yates
e-mail: andrew@ir.cs.georgetown.edu

N. Goharian
e-mail: nazli@ir.cs.georgetown.edu

O. Frieder
e-mail: ophir@ir.cs.georgetown.edu

J. Parker
Department of Emergency Medicine, Johns Hopkins University, Baltimore, USA

1 Introduction

During Twitter's initial public offering (IPO) in September of 2013 it filed a document (SEC 2013) with the US Securities and Exchange Commission claiming it had "200,000,000+ monthly active users" and processed "500,000,000+ tweets per day." The perceived value of these users and the data they generate resulted in an IPO that raised over 2.1 billion dollars. Some of this value is undoubtedly derived from information that can be gleaned from this daily deluge of tweets. Extracting this information from the continuous flood of tweets requires data organization.

Early adopters of Twitter also recognized the need for additional organization. On August 23rd of 2007 Chris Messina, whose Twitter username is factoryjoe, tweeted:

> "how do you feel about using # (pound) for groups. As in #barcamp [msg]?"

This tweet is credited as the first use of the # character as a hashtag on Twitter (Parker 2011). As the use of hashtags grew, the meaning of a hashtag evolved from indicating a group of people to indicating a topic and/or concept. In 2010 Twitter's front page began publishing a list of "Trending Topics" containing a list of hashtags that were being used with increasing frequency (Bowman 2010). At

Published online: 05 March 2015

that point, hashtags became an integral part of how Twitter was used on a daily basis.

Given their simplicity, hashtags are quite effective at organizing information, however, we cannot rely on them if we wish to collect all tweets pertaining to a particular topic; they are simply not required to publish a tweet. Moreover, many users may not be aware that their tweet is relevant to a topic we are interested in. For example, a user who tweets "ug terrible headache" early Saturday morning may not know that this tweet could have been tagged with a hashtag like #bingeDrinking or #riskyBehavior.

Our goal focuses on enabling "trending topic" detection for health-related topics without specifying a topic of interest in advance. In contrast to many prior efforts, we do not specify a particular health topic and then process tweets (or other social media messages) to investigate the rate at which that particular health topic is discussed over time. Our method is more general because it automatically generates many "interest curves" that correspond to many different health-related topics. Trends detected in these interest curves are referred to as health-related hypotheses because the input data (a large corpus of tweets) do not contain direct observations about the health of any single person, and the trend being detected is a shift in the vocabulary people use when they author tweets. Specifically, an output interest curve depicts increases in the prevalence of word sets previously used to discuss a particular health-related topic.

A high-level illustration of the framework presented here is shown in Fig. 1. The framework accepts a large corpus of time-stamped social media messages as input (tweets are used in this case). We filter the raw input corpus retaining only messages that are somewhat likely to be relevant to our broad topic of interest (i.e., health-related messages). Next, we partition the filtered corpus by time and find frequent word sets in each mini-corpus. After identifying a frequent word set, we create a time-series plot of that word set's prevalence in each time-based mini-corpus. This time series is used to determine if/when a frequent word set is trending. Once trending frequent word sets are identified, we connect trending frequent word sets with multiple topics (e.g., {runny, nose} → face, anatomy, sinus cavity, influenza …). For each connection between a trending word set and a topic, we create a (topic, timeWhenTrending) pair. We aggregate across all of these pairs to obtain a time series for each topic that depicts when that topic was associated with trending word sets. We refer to this time series as an interest curve. Finally, we filter out topics that are unimportant and return interest curves that can be further analyzed.

The backbone of this framework is the combination of frequent word set mining and information retrieval methods. Used together, these methods perform a powerful transformation on a signal deeply embedded in the "stream of raw tweets" domain to a simple signal in the "health related topics" domain. The transformation is performed using frequent item-set mining techniques, time-series analysis, and information retrieval methods.

In summary, the hypothesis generation framework presented here is:

- Designed to generate multiple "interest curves" for health-related topics and/or concepts. Each of these curves can be analyzed to extract hypotheses for future study and evaluation.

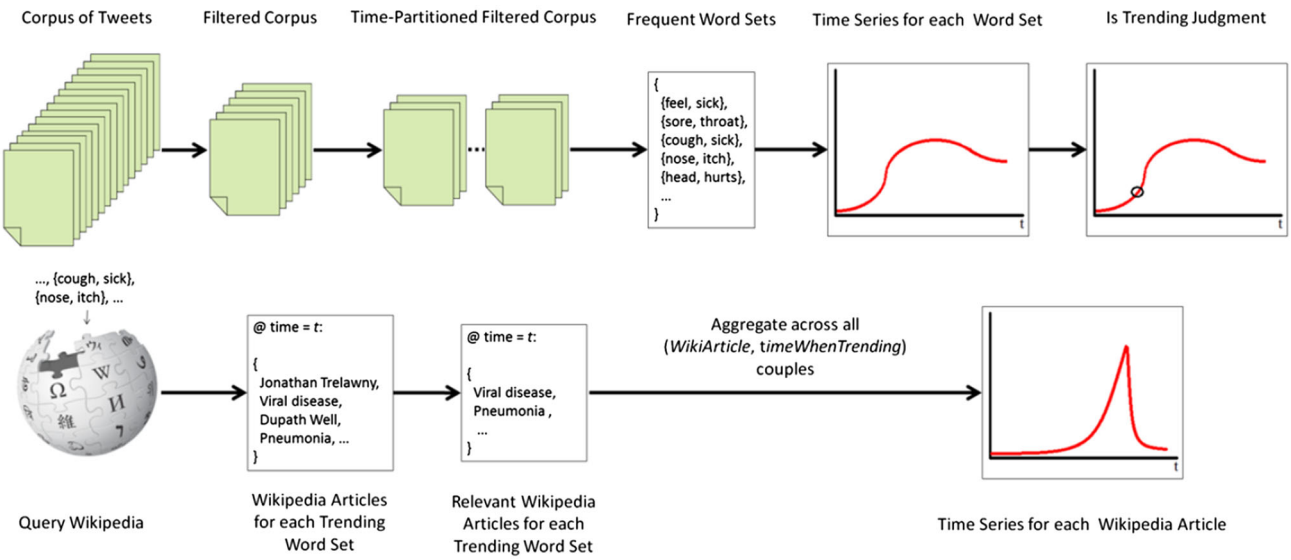

Fig. 1 A high-level view of the core algorithm

- Based on a transformation that converts a signal deeply embedded in the "stream of raw tweets" domain to a signal in the "health related topic" domain.
- In contrast to much prior art, is not designed to detect a single previously specified topic or concept of interest.
- Flexible and permits usage in alternate domains by replacing our health-specific filters with alternate filters.
- Built using mature, open-source resources thus making it a simple, efficient, and low-cost system which is ideal for practical use.

The source code used to generate the results within this paper is available at: https://github.com/Georgetown-IR-Lab. However, due to file size limitations, the corpus of 2 billion tweets, the corpus of 1.6 million health-related tweets, and the 10 GB compressed Wikipedia "dump file" cannot be obtained via these links.

2 Motivation

Twitter has been shown to be a reliable source for tracking public opinion about topics that range from political issues (O'Connor et al. 2010; Tumasjan et al. 2010), to natural disasters (Sakaki et al. 2010) and brand sentiments (Jansen et al. 2009). Even personal health is actively discussed in social media. People with chronic diseases like cancer are using social media to discuss their health, share stories, and provide peer-to-peer help with increasing frequency (Chou et al. 2009). A recent survey revealed that 26 % of "online" US adults discussed their health issues online in the past 12 months, and 42 % of them use social media to post or seek information about health conditions (Business Wire 2012). These facts suggest that social media content reflects, at least in part, public health conditions and can potentially serve as a source for public health surveillance systems.

Traditionally, public health surveillance systems are managed by professional health institutions like the Centers for Disease Control and Prevention (CDC) and the European Centre for Disease Prevention and Control (ECDC). Institution like these expend considerable effort collecting and analyzing clinical data to publish weekly surveillance data, as well as warnings of epidemic outbreaks. Importantly, both of these products usually reflect a one-to-two week reporting delay (Ginsberg et al. 2008).

A somewhat recent non-traditional approach is embodied in the Global Public Health Intelligence Network (GPHIN). GPHIN captures epidemic outbreaks by monitoring global media sources (essentially news websites) and at one point supplied approximately 40 % of the World Health Organization's (WHO) early warnings (Mykhalovskiy et al. 2006). Since GPHIN's success is largely attributed to the incorporation of comprehensive information from global news websites (Mykhalovskiy et al. 2006), it is once again reasonable to infer that social media—and Twitter in particular—could enable the creation of low-cost (as compared to traditional surveillance approaches) public health indicators and surveillance systems.

Healthmap (Freifeld et al. 2008), a system for processing and aggregating information on disease outbreaks from a wide range of electronic sources, has also demonstrated the value of combining outbreak information from traditional expert sources and less authoritative sources such as news media. This again suggests that information from less authoritative sources, such as Twitter, can be used to augment information from expert sources.

Due to this clear potential, there have been several Twitter-based public health monitoring approaches (Aramaki et al. 2011; Corley et al. 2009; Culotta 2010; Ginsberg et al. 2008; Jamison-Powell et al.2012; Lampos and Cristianini 2012; Paul and Dredze 2012; Paul and Girju 2010; Wenerstrom et al. 2012). However, most of these efforts focus on detecting a pre-established health condition (e.g., influenza or insomnia) and also assume that the condition is present. In contrast, we propose a general framework for identifying health conditions that may be emerging without relying on prior knowledge of (or assumption regarding) a condition's existence. In other words, while other approaches address questions like "Is such-and-such illness an increasingly prevalent health condition?", we address the more general question "What health conditions seem to be increasingly prevalent?" (with an answer that may include, but is not limited to the condition presupposed by other approaches).

The ideal long-term goal of creating an automated general purpose public health trend detector is to make a concrete impact on health outcomes. Achieving this goal requires an efficient detection method so that planners and decision makers can get "in front of" a health crisis. There is a vast body of disease simulation literature that seeks to clarify public health decisions like "Should schools be closed?" (Brown et al. 2011) and "Should international travel restrictions be put in place?" (Epstein et al. 2007). Furthermore, simulation techniques are now powerful enough (Parker and Epstein 2011) that better public health decisions could be made with less angst if accurate and timely disease surveillance data were available for thorough simulated cost/benefit analysis. Note, however, that regardless of which disease surveillance methods are used, there will always be public health officials vetting and inspecting the surveillance data.

3 Related work

Multiple efforts focus on tracking epidemics with tweets. Most of these efforts target the detection of influenza. Early

work by Corley et al. (2009) directly correlate occurrence of text which contain manually picked influenza-related words with official data (i.e., correlating the occurrences of the blog posts containing "influenza" or "flu" with Influenza Like Illness (ILI) rates). Similarly, Ginsberg et al. (2008) show a correlation between the occurrence of search queries containing flu-related words and ILI rates, and McIver and Brownstein (2014) show a correlation between views of flu-related Wikipedia pages and ILI rates.

To reduce human involvement and explore the entire feature space, Culotta (2010) proposed a model for automatically selecting textual features useful for labeling tweets as health related, which are later employed in tracking ILI rates. An improved version by Lampos and Cristianini (2012) employs a bootstrapping algorithm to extract a set of textual features from a tweet corpus using different feature selection principles. Additionally, Aramaki et al. (2011) train a support vector machine to label tweets as flu related or flu unrelated, and then evaluate the correlation of flu rates and flu-related tweets.

Rather than correlating the occurrence of flu rates and flu-related tweets, Wenerstorm et al. (2012) proposed a summarization method for flu-related tweets. According to their method, each flu-related tweet is represented with a vector of probabilities, each component of which corresponds to the tweet's probability of coming from a particular topic. A pairwise similarity value between tweets is derived from tweets' probability vectors, based on which tweets are clustered in a hierarchical or an agglomerative way. Tweets within the same cluster are ranked using closeness centrality, and common words of top ranking tweets summarize the cluster. When a Twitter monitoring system based on counting flu-related tweets signals, a flu outbreaks alarm, the summarization system allows health officials to quickly verify outbreak alarms.

Twitter is employed to study and monitor other ailments and health concerns in addition to influenza. Jamison-Powell et al. (2012) conducted a thematic analysis of insomnia-related tweets to reveal the degree to which people are using Twitter to discuss their mental health and how exactly they are doing it. Nakhasi et al. (2012) investigated patient perspectives on medical errors by exploring Twitter messages for self-reported adverse medical events. Diaz-Aviles et al. (2012) presented a personalized tweet ranking algorithm that could provide users a personalized, short list of tweets based on his or her own tweet context. Zhu and Goharian (2013) also report personalized Twitter information.

While all the above research targets a specific illness or health concern, a system capable of monitoring multiple ailments and health concerns are of more practical use. One appealing class of techniques for extracting information on multiple health conditions is probabilistic topic modeling. Techniques within this class include Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP), and Non Negative Matrix Factorization (NNMF). These methods model the association of terms with hidden topics, and view documents as a multinomial mixture of hidden topics (Chang et al. 2009). The topics discovered from these topic-modeling approaches need to be manually evaluated and often represent a mixture of topics (Blei et al. 2003) as opposed to one concept (e.g., a specific ailment).

In an effort to prevent this undesirable concept mixing, Paul and Dredze (2012) proposed the Ailment Topic Aspect Model (ATAM) that isolates various ailments within a corpus of tweets. Although ATAM is derived from LDA, it can output more coherent ailments that can be easily hand labeled with apt general titles such as obesity, respiratory, and dental. Both LDA and ATAM contain parameters that require tuning. ATAM's tuning relied on a specially focused corpus of health-related tweets they constructed. Although the method described herein differs vastly from the ATAM approach, we use their corpus of health-related tweets in our evaluation.

The topic detection and tracking (TDT) literature is similar in spirit to our method. Work in this area includes online event detection (Allan et al. 1998) and detecting "bursts" of activity in streaming text data (Kleinberg 2003). Koike et al. (2013) adapt Kleinberg's method to compare and contrast information with news streams and Twitter streams. These methods are more general than the health-related methods discussed above, but they still fall into the class of methods that identify topics first and then look changes in time-series data.

Finally, in Parker et al. (2013) we informally presented the idea of hypothesis generation from social media sources and provided limited experimental evaluations. Herein, we extend this effort by generalizing the hypothesis generation framework and furthering the experimental evaluations.

4 Twitter corpus

ATAM, as previously described, was trained using a corpus of 1.6 million health-related tweets that was culled from a much larger corpus of 2 billion tweets. The larger corpus of tweets was collected in part by O'Connor et al. (2010) and contains tweets from May 2009 to October 2010.

The work presented herein uses the corpus of health-related tweets. This narrowly focused corpus was created by removing 99.92 % of the content from the larger twitter corpus using a multiple pass filter. The first filtering pass removed tweets that did not contain at least one of 20,000 key words and phrases related to illnesses/diseases,

symptoms, and treatments that were scraped from wrongdiagnosis.com and mtworld.com (the exact links are no longer valid, but the resulting lists of words and phrases is included with our source code). The second pass removed re-tweets and tweets containing URLs. The final and arguably most important filtering operation applied a custom built SVM classifier. The SVM classifier was trained using data collected from Mechanical Turk and was designed to favor high precision over high recall. It is worth noting that this SVM classifier removed non-English tweets. A more detailed description of the filtering process and the SVM classifier in particular, can be found in Paul and Dredze (2012).

5 The framework

The goal of our framework is to enable automatically detecting possible emerging public health concerns using Twitter. We want to do this without designating a priori which public health concern(s) is (are) most important. In other words, we want to interact with our system to discover possible emerging public health concerns (e.g., "Question: What illnesses seem to be occurring more frequently lately? Answer: Flu") rather than retrieving feedback regarding a single user-specified health concern (e.g., "Question: Is flu occurring more frequently lately? Answer: Yes").

Our framework is based on a core assumption that people will describe the chief complaint (i.e., primary symptoms) of an illness on Twitter (which for some conditions, like sexually transmitted diseases, is unlikely to be true). Our framework is designed to find illnesses and other medical concepts described using the same vocabulary people use when authoring tweets on Twitter.

To provide the desired capability, our framework leverages three mature open-source resources: Mahout, Lucene, and Wikipedia. The parallel FP-Growth (Li et al. 2008) implementation in Mahout is used to find frequent word sets. Traditional information retrieval searches are performed to connect word sets with Wikipedia articles. These searches are performed programmatically using a Lucene index containing the complete database of Wikipedia articles. A high-level view of the algorithm is shown in Fig. 1; a detailed description of each step is described in the following subsections.

The actions of this framework can be seen as a transformation from one domain to another. Just like the Fourier Transform and the FFT transform a signal from the time domain to the frequency domain of our framework transforms a signal deeply embedded in a stream of raw tweets to a signal describing a health-related topic.

5.1 Filtering the initial corpus

The first step is to filter the raw corpus of tweets and retain only those tweets that are reasonably likely to be relevant to the broad topic of interest. This filtering step is beneficial because it significantly reduces the number of tweets that must be processed in subsequent steps. We used the corpus of 1.6 million health-related tweets Paul and Dredze (2012) created by filtering the original corpus of 2 Billion tweets.

Strictly speaking, a filtering step may not be necessary. The trends detected may be detectable even when processing the entire corpus of 2 Billion tweets. We did not investigate this possibility. However, in the future, we plan to investigate the feasibility of using a filtering method based solely on simple operations like key word matching. It is our hope that similar trends can be detected without relying on a filtering method that requires training (like the SVM filter used to obtain the corpus of 1.6 million health-related tweets).

5.2 Partitioning the corpus by time

The next step towards implementing our framework is to partition the filtered corpus into multiple mini-corpuses based on time. We present results from using two different time-based partitioning methods. We divided the corpus by the month in which each health-related tweet was published, as well as by the week it was published. Predictably, the variability in which word sets are considered "frequent" and "trending" is higher when partitioning the corpus by week of publication. Note that newer datasets may exhibit less variability as the number of tweets published per day has risen significantly, since our data were collected.

5.3 Finding frequent word sets

Before frequent word sets can be found within a mini-corpus, the tweets within it must be standardized. The raw text of each tweet is standardized using the following operations:

- Punctuation characters are replaced with spaces.
- All text is converted to lowercase.
- The text is tokenized.
- Stop words are removed.
- Duplicate tokens are removed.

After standardization, each tweet is treated as a set of words that can be analyzed using off-the-shelf association rule mining techniques. In particular, we use the parallel FP-Growth implementation within Apache's data mining library Mahout to find the frequent word sets within each mini-corpus. We opt to use FP-growth over similar

methods like the Apiori algorithm (Agrawal and Srikant 1994) or the ECLAT algorithm (Zaki 2000) because FP-growth is the fastest of these algorithms and, equally importantly, a high quality implementation of FP-growth is available from the open-source community. While using FP-growth we vary the minimum support used when mining each mini-corpus to ensure that the conceptual definition of "frequent" remains constant across the various mini-corpora. The minimum support is always set to the smallest integer n such that n is at least 0.1 % of the tweets within that particular mini-corpus. Consequently, any set of words that do not reach this threshold for a particular month or week will not be listed in the collection of frequent word sets extracted from the corresponding mini-corpus.

Our framework allows for word sets to be identified using methods other than frequent pattern mining (i.e., parallel FP-Growth). For example, a system for identifying mentions of adverse drug reactions (ADRs), ADRTrace by Yates and Goharian (2013), was used to identify word sets related to ADRs. This is a different approach than finding word sets using frequent pattern mining; ADRTrace uses a dictionary (created using synonym discovery methods from Yates et al. 2014) and syntactic patterns to identify mentions of ADRs. This domain-specific word set identification technique is being developed for detecting trends related to pharmacovigilance (e.g., an ADR might begin to trend after a drug release). These drug-related trends can then be used by a pharmacovigilance surveillance system (e.g., Burger et al. 2013) to identify drugs that may be causing unexpected ADRs.

5.4 Creating time series for word sets

After mining a mini-corpus, we have a collection of frequent word sets like {{flu, sick}, {headache, feel}, {hurts, sick, throat}, {feeling, stomach} …}. For each frequent word set, we build a time series that shows how prevalent that particular word set is in each mini-corpus. An example is shown in Fig. 2. These time series are used to determine which word sets have recently seen a significant increase in prevalence, that is, which word sets are trending.

5.5 Make "is trending" decisions for word sets

We cannot detect potentially interesting trends in Twitter data merely by observing that some word sets are common. For example, the word set {feel, sick} is the most prevalent word set in almost every partition of our dataset (because stop words have been removed and the SVM filter was designed to find a specific flavor of tweet). Therefore, we examine a frequent word set's time series to determine if and when the word set becomes significantly more prevalent than it was in the recent past. In other words, we determine when the word set "is trending."

To make the "is trending" decision, we use a simple rule based on the growth rate of a word set's prevalence at time t (see FreqTimeSeries.java in the source code for

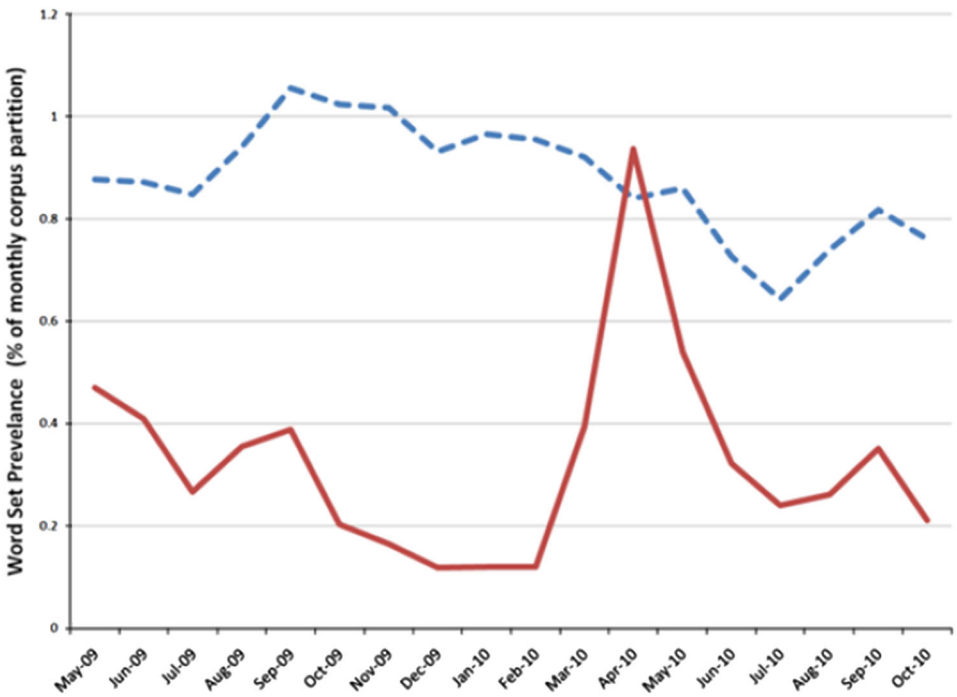

Fig. 2 Prevalence of two frequent word sets by month: *solid line* "allergies feel", *dashed line* "feel sick". Notice, the *dashed line* is almost always more prevalent than the *solid line*; however, the *dashed line* does not appear to trend at any point in time

more). Basing the decision on prevalence (i.e., the fraction of tweets in a time-partition containing a particular word set), as opposed to raw counts, is helpful because the dataset was collected, while the Twitter usage was increasing significantly month by month.

It is important to note that making an "is trending" decision from a time series is a complex research topic on its own. Some notable methods in this area include control chart (Shewhart 1931), exponentially weighted moving average (Roberts 1959), CUSUM (Page 1954), Box-Jenkins models (Box et al. 1970), and Kalman Filters (Kalman and Bucy 1961). To maximize simplicity, we do not incorporate one of these more complex methods when analyzing the time series for a particular word set. We feel this simplification is acceptable because the ultimate goal of detecting a trend at the health topic or health condition level requires detecting when several trends are occurring simultaneously across different, but topically related, word sets. Therefore, as long as a "is trending" rule used for word sets is accurate enough to detect the majority of co-occurring trends (across multiple word sets), it does not matter if the rule fails to detect a trend in a small minority of cases because the broader pattern is not lost.

5.6 Query wikipedia

We use Wikipedia to associate trending frequent word sets with the articles (i.e., topics) found in Wikipedia. Wikipedia is a good choice for this role due to its wide coverage and the fact that it is written in layman's English (closely resembling the tweets considered). Once a list of articles is obtained, we filter out topics that are not pertinent to public health.

Using Lucene, we indexed the complete Wikipedia compressed archive (available at: http://en.wikipedia.org/wiki/Wikipedia:Database_download). Before each Wikipedia article is indexed, we parse and store the article's introduction and info boxes if they exist. We explicitly store these fields because they are used to determine which Wikipedia articles may be relevant to public health. The index is built using the Standard Analyzer from Lucene version 3.5.

We push every trending frequent word set as queries through our search system. These queries return Wikipedia articles that were deemed relevant to the input query (i.e., "runny nose" → Rhinorrhea …). For each query result, we create a (*WikiArticle*, *timeWhenTrending*) pair. These pairs are later aggregated to determine which Wikipedia articles are repeatedly associated with trending word sets (that may or may not trend at the same time).

5.7 Filtering wikipedia results

Many Wikipedia articles returned during the previous step have no connection to public health concerns. For example, articles about famous people are frequently returned if that person once showed symptoms that could be described using the search terms (e.g., "fever flu" → Barry Bonnell the baseball player). We convert every frequent word set to a query and filter results. Filtering is a necessary step as it is difficult to programmatically determine a priori which word sets will generate health-related topics. Two filtering methods were considered. The first method only returns Wikipedia articles containing ICD codes (see below). The second method returns not only Wikipedia articles containing ICD codes but also articles with introductions that contain a large proportion of medically related words.

5.8 Precision filter

The first filtering method used to differentiate health-related Wikipedia articles from non-health-related articles is based on the presence (or lack thereof) of an ICD code within the article. The ICD coding system is an international standard classification system that has been used extensively to encourage inter-operability of medical and insurance computer systems. The 10th revision of ICD, ICD-10, contains over 14,440 different codes distributed across different sub-classes like diseases and medical procedures. Figure 3 shows a typical Wikipedia article that has an info box containing an ICD code. Finding an ICD code within an info box is a strong indicator that the article is medically relevant. The strength of this required indicator ensures that the set of articles that pass this filter will have a significant health aspect to them.

5.9 Recall filter

The second filtering method, we consider is more inclusive; thus, its recall is higher than the prior filter. This second filtering method accepts every article that the precision filter accepts, as well as articles containing "medically relevant" introductions.

When we use the term "introduction" we must be careful because Wikipedia articles do not have an officially labeled Introduction section. However, Wikipedia articles generally do have labeled sections. The "Sore Throat" article, a portion of which is shown in Fig. 3, has the following five sections: Definition, Differential Diagnosis, Treatment, Epidemiology, and References. We classify any text that comes before the first labeled section as the introduction of that article. We do not include info boxes as part of the introduction even though the text that defines

Fig. 3 A snippet from a typical Wikipedia article: The introduction and info *box* are enclosed in *rectangles*. The ICD codes are *circled*

them (in Wikipedia's markup language) appears before the first labeled section.

Once an article's introduction is isolated, we analyze the introduction to determine if it is "medically relevant." To make this determination, we:

- Tokenize the introduction
- Remove stop words
- Count the tokens and medical tokens
- If: token count ≤ 10

  Then: return "is not medical"

- If: numMedicalTokens/numTokens $\geq .75$

  Then: return "is medical"
  Else: Return "is not medical"

The steps shown above require the ability to determine if an individual token is medical. We make this determination by searching for the token in Stedman's Medical Dictionary available online at http://www.medilexicon.com/medicaldictionary.php.

5.10 Aggregating trends

This framework requires two complex events to occur before it will generate a signal that suggests a particular health condition is occurring more frequently. The first of these complex events is that multiple trending word sets must be associated with the same Wikipedia article. For example, {sore, throat}, {nose, runny}, and {cough, nose} all list the "Common Cold" article within their respective query results. The second required event is that those trending word sets must all trend at the same time (or nearly so). When both of these events occur it is unlikely that their occurrence is due to chance. The process of aggregating the (*WikiArticle*, *timeWhenTrending*) pairs by common, Wikipedia article produces a time series depicting the number of times these complex events occurred simultaneously. The unit on the y-axis of these time series is "number of trending word set associated with Wikipedia article XYZ," while time is on the x-axis. We call each of these time series an interest curve.

5.11 A flexible approach

This framework is built by chaining together results obtained from multiple independent sub-steps. Many of these sub-steps can be implemented differently without dramatically changing the output of the system as a whole. For example, the initial filtering method may not need to be as precise, the "is trending" algorithm could be one of many other methods (see Sect. 5.5), the Lucene search system need not use the default configuration, and the technique that identifies health-related Wikipedia articles could reflect standard document classification techniques. We do not examine this plethora of possibilities because

optimizing design choices that will have subtle impacts should occur closer to commercialization or operationalization in a specific domain.

6 Results

Our results confirm that seasonal increases in common health conditions are indeed detectable without using search strategies explicitly customized to detect those specific health conditions. In particular, we observe (among other things) allergy season, flu season, and even a small uptick in summertime ice-cream headaches (i.e., "brain freeze") using one general purpose algorithm. Our results also illustrate that our methodology is likely to highlight multiple medical conditions with similar symptoms as opposed to highlighting just one or two conditions that could be considered the "best response" for a particular trending word set. For example, interest curves for several different types of headaches are generated as are interest curves for multiple respiratory ailments like influenza, the common cold, cough, and acute bronchitis.

6.1 System output and operationalization

The system from Sect. 5 generates numerous time-series curves as output. Each of these interest curves is associated with a different health-related topic or condition and can be processed further (if desired). For example, one of the time-series analysis methods mentioned in Sect. 5.5 could be applied to produce a binary "anomaly/no anomaly" or "trend/no trend" style output. Another alternative is to produce an output that returns a tiered warning level like green, yellow, or red.

This transition from a time series to a binary or tiered output is non-trivial. Decisions made during this transition will directly impact the overall sensitivity and specificity of the system. These decisions may also vary depending on the health-related topic or condition at issue. There is no reason, other than simplicity, for every topic/condition's time series to be analyzed using the same level or risk-tolerance. This complexity means deploying a system that delivers binary or tiered output requires a user to carefully consider his or her risk-tolerance and willingness to trade false positives for false negatives.

6.2 Influenza

The curve in Fig. 4 shows the number of times a trending word set is associated with the "Influenza" Wikipedia when the corpus of tweets is partitioned by month. By comparing the results in Fig. 4 with true influenza incidence shown in Fig. 5, we can see that our framework produces the weakest signal (i.e., the smallest values) when the slope of the true incidence is negative. The curve in Fig. 4 also produces its strongest signal when the slope of the true incidence curve is strongly positive (in September and October of 2009). The beginning of the mild 2010 flu season also coincides with an uptick in Fig. 4. It is important to point out that the curves in Figs. 4 and 5 should

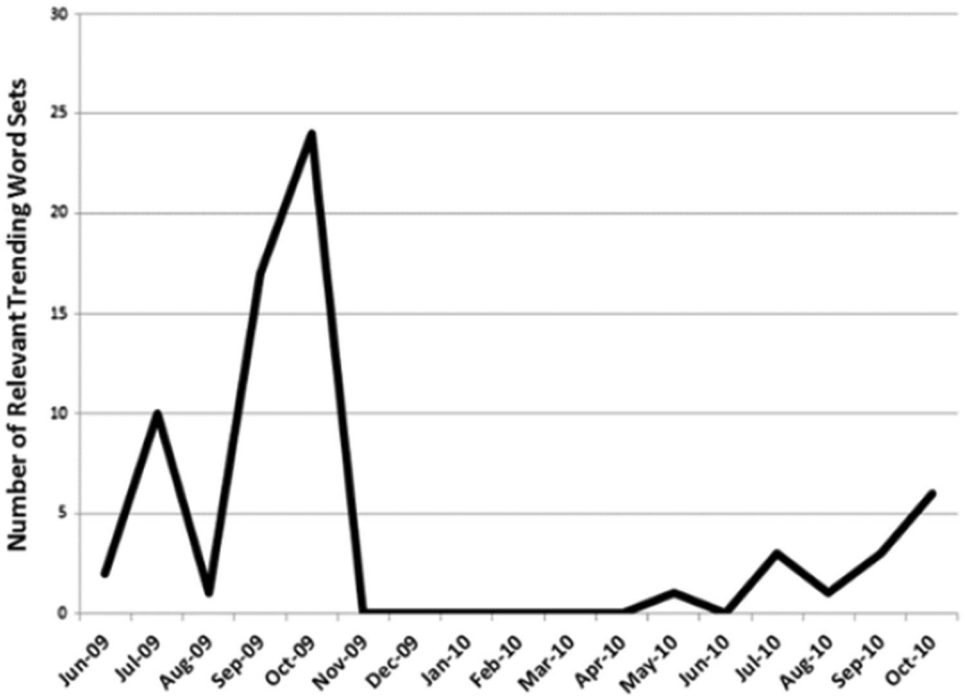

Fig. 4 The number of trending word sets associated with the "Influenza" Wikipedia article when using monthly time partitioning

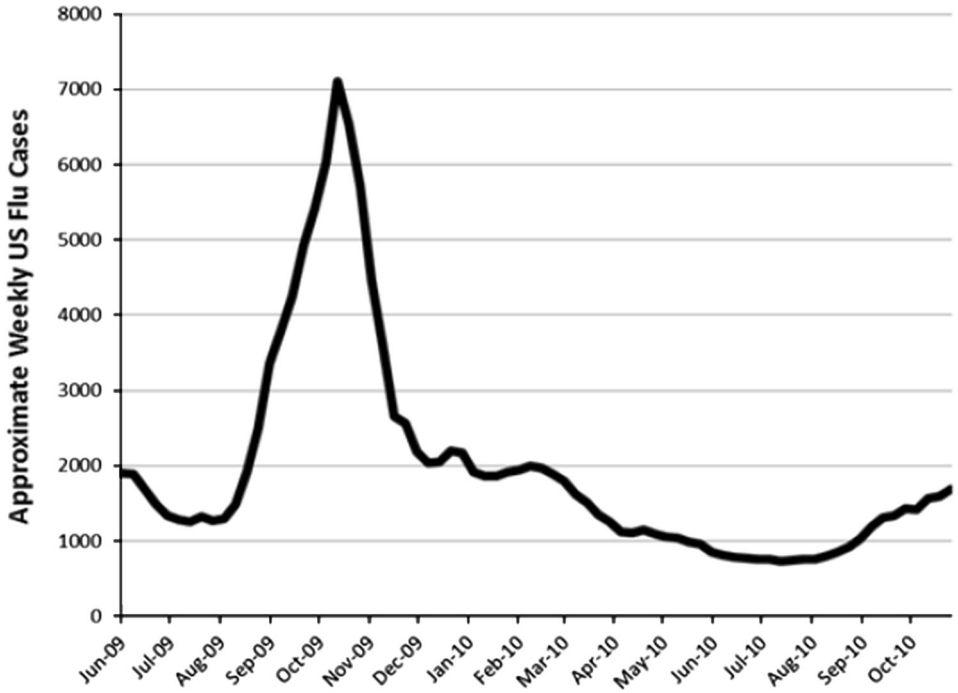

Fig. 5 Approximate weekly flu cases in the United States from June 09 to October 2,010 (Flu Trends)

not be proportional to one another or linearly related. However, the flu incidence data shown in Fig. 5 can be transformed to enable a meaningful comparison with the data shown in Fig. 4. To enable this comparison, we compute the derivative of the flu incidence data and set all negative values to 0 (i.e., we compute max(0, fluIncidence$'(t)$)). We can then compute the correlation between the number of trending word sets associated with the "Influenza" Wikipedia article and the transformed flu incidence data. This correlation ranges from 0.763 when no smoothing is applied to 0.892 with smoothing.

In an ideal world, any non-zero entry in Fig. 4 curve would indicate real world influenza cases were indeed growing in number. However, this is not the case. The moderately strong detection signal seen in July of 2009 (when "Influenza" was associated with 10 trending word sets) does not correspond to a simultaneous increase in US flu cases. We attribute this data point to the notable increase in flu interest that occurred after the WHO raised the worldwide pandemic alert level to Phase 6 on June 11, 2009. It is possible that much of the lag from June 11 to July can be accounted for by the reporting delay for official CDC flu incidence numbers which typically required one-to-two weeks to gather, tabulate, and publish.

We duplicated much of the prior analysis except we partitioned our dataset by week of publication instead of month of publication. Figure 6 depicts this similar treatment. It is worth noting that the peaks in Fig. 4 correspond to quick but strong pulses (around July of 2009) or contiguous periods of sustained activity (September and October of 2009).

It should also be noted that the comparison between the curves in Figs. 4 and 5 is subject to one small caveat. Our corpus of 1.6 million tweets was not explicitly filtered to contain only the US-based tweets. However, we do not believe this is a significant problem because the SVM classifier used to filter the raw corpus was designed to discard non-English tweets.

6.3 Precision filter vs. recall filter

In Sect. 4, we mention that two different filters are used to separate health-related Wikipedia articles from non-health-related articles. We believe the precision focused filter that requires an ICD code to be within the article, is preferable to the recall focused filter which accepts either an ICD code mention or medically related terms in the introduction. The recall focused filter allows a few obviously non-health-related articles through but the majority of the additional articles merely define a body part or system (e.g., Mucous, Nasal cartilages, Cough reflex). Although the identification of a body part or system does provide additional information, it fails to further identify a general trended condition. Since we aim to identify a general health diagnosis, we prefer the precision focused filter over the recall focused filter.

6.4 Confounding by symptoms and syntax

Our methodology produces curves for 11 different articles related to one respiratory ailment or another. It also

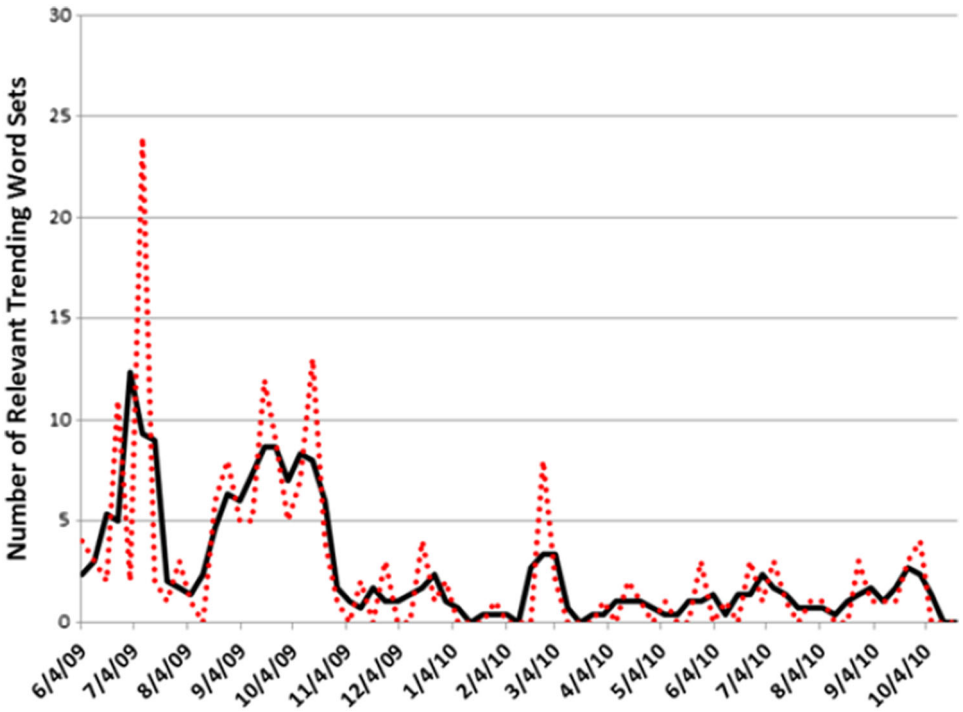

Fig. 6 The *number* of trending word sets associated with the "influenza" Wikipedia article when using weekly time partitioning (*dotted* raw data, *line* smoothed data)

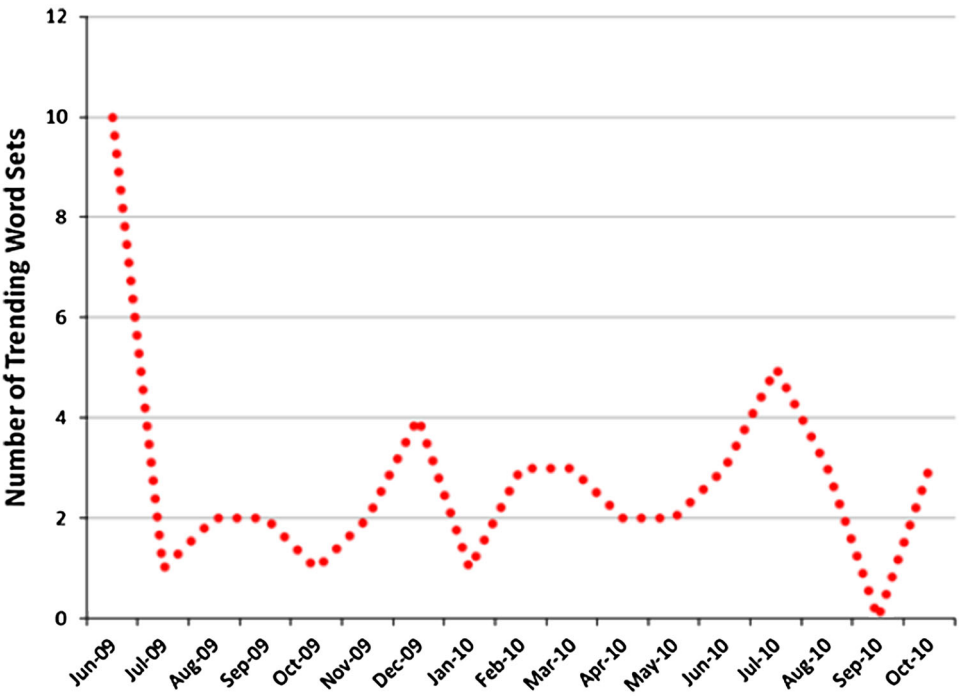

Fig. 7 The *number* of trending word sets associated with the "Ice-cream Headache" Wikipedia article

produces curves for 12 different Wikipedia articles about headaches and migraines. The interesting difference between these two groups is that the existence of each "family" is driven by markedly different phenomena. The group of respiratory results is created by tweets describing symptoms. For example, "runny nose" and "sore throat" both map to multiple respiratory conditions when those word sets are trending. The batch of headache results is driven by the two different meanings of the word headache: physical pain (e.g., "I bumped my head and now I have a headache") and annoyance (e.g., "My computer crashed—what a headache"). As a result of these disparate drivers,

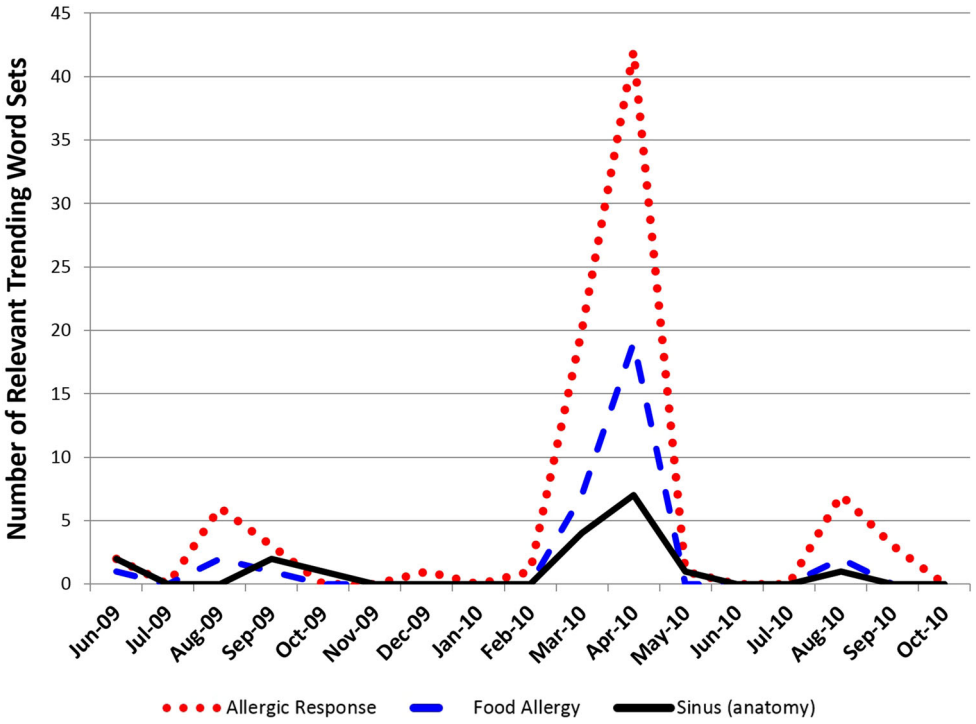

Fig. 8 The *number* of trending word sets associated with the "allergic response" (*dotted*), "food allergy" (*dashed*), and "sinus" (*solid*) Wikipedia articles when using monthly time partitioning

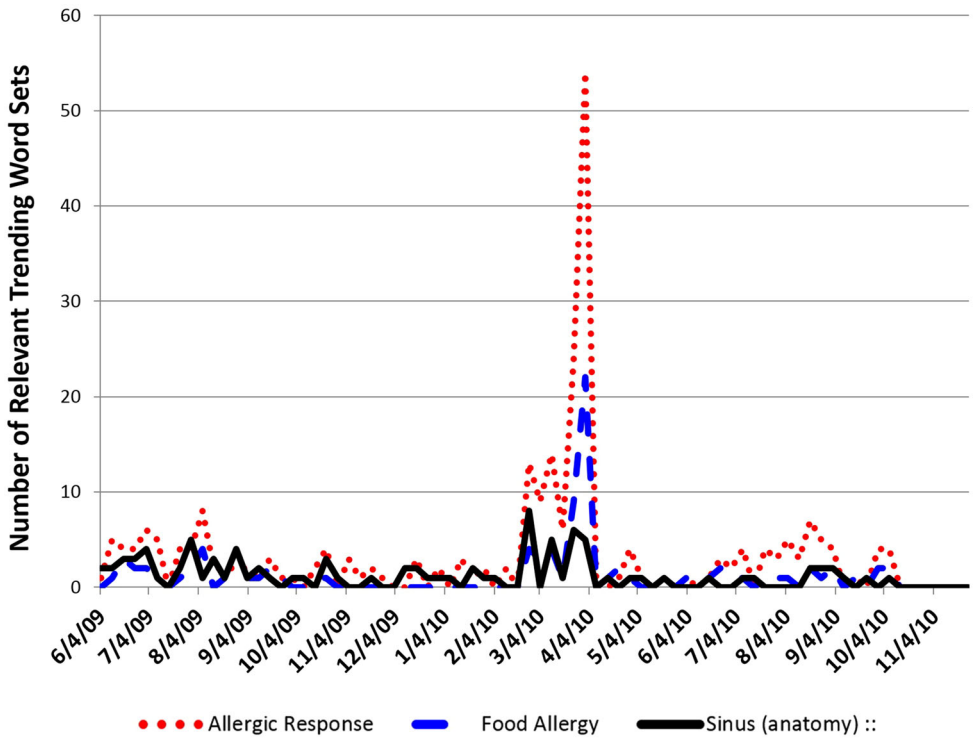

Fig. 9 The *number* of trending word sets associated with the "allergic response" (*dotted*), "food allergy" (*dashed*), and "sinus" (*solid*) Wikipedia articles when using weekly time partitioning

the signals associated with the family of respiratory results have a much better cohesion than the signals associated with the family of headache results.

Although the batch of headache results is confounded by the colloquial use of the word "*headache,*" some promising news within that collection exists. The "Ice-cream headache" article is associated with far fewer trending word sets than almost all other headache-related articles like "Vascular headache" and "Tension headache" (the retinal migraine article is the sole exception). This

reduced signal occurs because most queries involving the colloquial use of the word headache are not linked to the "Ice-cream headache" article because it is pushed too far down the Lucene search results by other more relevant headache articles. This is good because the signals associated with the word sets {eating, headache}, {headache, ice}, and {cream, headache, ice} (among others) are not drowned out by the multitude of signals emitted by the colloquial use of the word headache.

Since the number of trending word sets associated with the Ice-cream headache article is not confounded by colloquial uses of the word "headache" the fact that this number peaks in June of 2009 and July of 2010 (which, at the time, was the hottest month on record in many places throughout the US) is less likely to be a quirk due to random chance. Figure 7 shows the number of trending word sets associated with the "Ice-cream Headache" Wikipedia article.

Another type of confounding comes from multiple Wikipedia articles getting highlighted due to word sets like: {allergies, asthma}, {allergies, lol}, and {allergies, eyes, itchy}. These three word sets (and many similar word sets) all trend during the early spring. From the word sets themselves and the time they trend, it is clear that the underlying condition is the pollen-related allergies that are prevalent during the spring. On a positive note, we produce curves for multiple seasonal allergy-related Wikipedia articles—two examples of which are shown in Figs. 8 and 9. The problem is that interest curves for multiple food allergies are included in this batch. It is possible that a medical synonyms set as in Yates and Goharian (2013) may prove useful when addressing the problems that common symptoms present.

6.5 Duplicate detection is helpful

It seems reasonable to assume that when real-world medical problems are trending—and those problems are discussed on Twitter with a somewhat unique vocabulary—then we might expect word sets containing words from that vocabulary to also trend. Notice, we use the plural "word sets" because we expect multiple different sets to trend due to combinatorics. For instance, if n words are highly likely to be used when a person is writing about the flu on Twitter, we can expect several combinations of these n words to trend at the same time. We can also expect many of these word combinations to trend when paired with words outside the unique vocabulary, e.g., {flu, hate}. This possibility of repeat detection is helpful because it enables a way to gage the strength of an observed trend. Topics that are flagged by many word sets are likely to be better hypotheses than topics that are flagged by only a few word sets.

6.6 Results and discussion

The results shown above are promising. Taken together they form a proof of principle. The framework generates curves for the well-known seasonal medical ailments of influenza and springtime allergies without any ailment-specific customization. These results were obtained, while a minimum support of 0.1 % word set prevalence was required (discussed in Sect. 5.3). We do not believe the minimum support must be set this high for the trend detection methodology to work. In other words, we do not believe this methodology is only good for detecting common conditions. In fact, we believe reducing the minimum support and searching for seasonal sports-related injuries would be a useful exercise. It would be promising if concussions and knee injuries were flagged as a trending health condition when the high school football season started because these injuries are common for football players and uncommon in the general population. Thus, if they are detectable then we would have good reason to believe that other somewhat rare health conditions could also be detected. It is likely that this exercise may require minor alterations to the "*is trending*" algorithm and/or the time partitioning choice.

Our framework does produce some interest curves that are likely to generate false positives when processed further. For example, the interest curve for the "Food Allergy" article has a large uptick in March and April of 2010, because that article contains many of the same words people use to discuss pollen allergies on Twitter.

This work was performed using a corpus of health-related tweets that were culled from a larger corpus using three filters. It is unclear if using the filtered dataset generates better results. Due to the absence of a strong intuition about which corpus would be best we opted to use the more manageable corpus of 1.6 million tweets as opposed to the corpus of 2 billion tweets. In the future, we want to investigate using a significantly less restrictive filter that still reduces the problem size several orders of magnitude.

7 Source code

The source code used to generate the results in this paper is available at: https://github.com/Georgetown-IR-Lab. It is

pure Java and contains classes that make processing large corpuses of tweets more manageable. Some of the highlights include:

- Classes to build and query Lucene search systems.
- Methods to process large collections of tweets using Mahout's FP-Growth implementation.
- Classes that read directly from compressed.gz files.
- Classes that facilitates bulk manipulation and analysis of tweet text.
- Methods to clean, stem and tokenize text.
- Methods to partition large collections of tweets by time of authorship.

8 Conclusion and future work

We demonstrated a single framework for detecting a multitude of public health trends which clearly identified the seasonal afflictions of influenza, allergies, and summertime ice-cream headaches. The framework is simple to implement and operates efficiently because it is built on top of mature algorithms from association rule mining and information retrieval. A combination of techniques from these fields and some time-series analysis is used to transform a signal deeply embedded in the "stream of raw tweets" domain to a signal in the "health related topics" domain.

We output interest curves for health-related topics because we use a filtered corpus and embedded ICD codes to filter Wikipedia search results. We could conceivably focus on a different topic of interest by changing the filters to suit the new topic.

We have two main future development goals: (1) We would like to run the framework on a larger scale to reduce the variance seen in the weekly time series and, hopefully, enable increasing the temporal resolution from months and weeks to days and (2) We would like to investigate using a method/resource besides ICD codes to filter out non-medically related trending topics. Using Wikipedia and ICD makes detecting previously known (and possibly common) ailments easy. However, using ICD may also prevent the detection of novel ailments and effortlessly adapting the framework to suit another domain. We suspect standard document classification or clustering methods from the information retrieval literature could replace our ICD-based methodology.

Acknowledgments This work was partially supported by: the US National Science Foundation through Grant CNS-1204347, the Models of Infectious Disease Agent Study (MIDAS), under Award Number U01GM070708 from the NIGMS, The Johns Hopkins Medical School DHS Center on Preparedness and Catastrophic Event Response (PACER), under Award Number N00014-06-1-0991 from the Office of Naval Research, and Joshua M. Epstein's NIH Director's Pioneer Award, Number DP1OD003874 from the Office of the Director, National Institutes of Health. Finally, we would like to thank Social Network Analysis and Mining for the invitation to deepen our paper from the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.

References

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of 20th international conference on very large data bases, VLDB, pp 487–499

Allan J, Papka R, Lavrenko V (1998) On-line new event detection and tracking. In: Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval. ACM

Aramaki E, Maskawa S, Morita M (2011) Twitter catches the flu: detecting influenza epidemics using twitter. In: Proceedings of the Conference on empirical methods in natural language processing, EMNLP, pp 1568–1576

Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. the. J Mach Learn Res 3:993–1022

Bowman D (2010) "Tweaking the Twitter homepage", The offical twitter blog, posted 30 Mar 2010. https://blog.twitter.com/2010/tweaking-twitter-homepage

Box G, Jenkins G, Reinsel G (1970) Time series analysis: forecasting and control. John Wiley & Sons

Brown ST, Tai JH, Bailey RR, Cooley PC, Wheaton WD, Potter MA, Voorhees RE, Lejeune M, Grefenstette JJ, Burke DS, McGlone SM, Lee BY (2011) Would school closure for the 2009 H1N1 influenza epidemic have been worth the cost?: a computational simulation of Pennsylvania. BMC Public Health 11(1):353

Burger EW, Federoff H, Frieder O, Goharian N, Yates A (2013) Social media communications networks and pharmacovigilance: SequelAE-2.0. In: Proceedings of the IEEE 15th international conference on e-health networking, applications and services, healthcom

Business Wire (2012) Twenty six percent of online adults discuss health information online; privacy cited as the biggest barrier to entry.http://www.businesswire.com/news/home/20121120005872/en

Chang J, Boyd-Graber JL, Gerrish S, Wang C, Blei DM (2009) Reading tea leaves: how humans interpret topic models. In: Proceedings of the 23rd annual conference on neural information processing systems, NIPS, pp 288–296

Chou W, Hunt Y, Beckjord E, Moser R, Hesse B (2009) Social media use in the United States: implications for health communication. J Med Internet Res, 11(4)

Corley C, Mikler A, Singh K, Cook D (2009) Monitoring influenza trends through mining social media. In Proceedings of the international conference on bioinformatics computational biology, ICBCB, pp 340–346

Culotta A (2010) Towards detecting influenza epidemics by analyzing twitter messages. In: Proceedings of the 1st workshop on social media analytics, pp 115–122

Diaz-Aviles E, Stewart A, Velasco E, Denecke K, Nejdl W (2012) Towards personalized learning to rank for epidemic intelligence based on social media streams. In: Proceedings of the 21st international conference companion on world wide web, WWW, pp 495–496

Epstein JM, Goedecke DM, Yu F, Morris RJ, Wagener DK et al (2007) Controlling pandemic flu: the value of international air

travel restrictions. PLoS ONE 2(5):e401. doi:10.1371/journal.pone.0000401
FluTrends. http://www.google.org/flutrends/us/#US
Freifeld CC, Mandla KD, Reis BY, Brownstein JS (2008) Health map: global infectious disease monitoring through automated classification and visualization of internet media reports. J Am Med Inform Assoc
Ginsberg J, Mohebbi M, Patel R, Brammer L, Smolinski M, Brilliant L (2008) Detecting influenza epidemics using search engine query data. Nature 457(7232):1012–1014
Jamison-Powell S, Linehan C, Daley L, Garbett A, Lawson S (2012) I can't get no sleep: discussing# insomnia on twitter. In: Proceedings of the ACM annual conference on human factors in computing systems, CHI, pp 1501–1510
Jansen B, Zhang M, Sobel K, Chowdury A (2009) Twitter power: tweets as electronic word of mouth. J Am Soc Inform Sci Technol 60(11):2169–2188
Kalman R, Bucy R (1961) New results in linear filtering and prediction theory. J Basic Eng 83(1):95–108
Kleinberg J (2003) Bursty and hierarchical structure in streams. Data Min Knowl Disc 7(4):373–397
Koike D, et al. (2013) Time series topic modeling and bursty topic detection of correlated news and twitter. In: Proc. 6th IJCNLP
Lampos V, Cristianini N (2012) Nowcasting events from the social web with statistical learning. ACM Trans Intell Syst Technol 3(4):72
Li H, Wang Y, Zhang D, Zhang M, Chang E (2008) PFP: parallel FP-growth for query recommendation. In: Proceedings of the ACM conference on recommender systems, pp 107–114
McIver DJ, Brownstein JS (2014) Wikipedia usage estimates prevalence of influenza-like illness in the United States in near real-time. PLoS Comput Biol 10(4):e1003581
Mykhalovskiy E, Weir L et al (2006) The global public health intelligence network and early warning outbreak detection: a Canadian contribution to global public health. Can J Public Health 97(1):42
Nakhasi A, Passarella R, Bell S, Paul M, Dredze M, Pronovost P (2012) Malpractice and malcontent: analyzing medical complaints in twitter. In: AAAI Fall Symposium Series
O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: linking text sentiment to public opinion time series. In: Proceedings of the 4th international conference on weblogs and social media, ICWSM
Page E (1954) Continuous inspection schemes. Biometrika 100–115
Parker A (2011) Twitter's Secret Handshake. The New York Times. Retrieved 26 Jul 2011. http://www.nytimes.com/2011/06/12/fashion/hashtags-a-new-way-for-tweets-cultural-studies.html?_r=2&pagewanted=all&
Parker J, Epstein JM (2011) A distributed platform for global-scale agent-based models of disease transmission. ACM Trans Model Comput Simul. 22(1) Article 2, p 25
Parker J, Wei Y, Yates A, Frieder O, Goharian N (2013) A framework for detecting public health trends with twitter. In: Proceedings of the international conference on advances in social networks analysis and mining
Paul M, Dredze M (2012) A model for mining public health topics from twitter. HEALTH 11:16–26
Paul MJ, Girju R (2010) A two-dimensional topic-spect model for discovering multi-faceted topics. In: Proceedings of the 24th AAAI conference on artificial intelligence
Roberts S (1959) Control chart tests based on geometric moving averages. Technometrics 1(3):239–250
Sakaki T, Okazaki M, Matsuo Y (2010) Earthquake shakes twitter users: real-time event detection by social sensors. In: Proceedings of the 19th international conference on world wide web, WWW, pp 851–860
Shewhart W (1931) Economic control of quality of manufactured product. vol 509. ASQ Quality Press
SEC Amendment 1 to Form S-1 Registration Statement, Twitter,Inc. EDGAR. October 15, 2013. Retrieved 8 Nov 2013. http://www.sec.gov/Archives/edgar/data/1418091/000119312513400028/d564001ds1a.htm
Tumasjan A, Sprenger TO, Sandner PG, Welpe IM (2010) Predicting elections with twitter: what 140 characters reveal about political sentiment. In: Proceedings of the 4th international conference on weblogs and social media, ICWSM
Twitter statistics. http://www.statisticbrain.com/twitter-statistics/
Twitter blogs: measuring tweets. http://blog.twitter.com/2010/02/measuring-tweets.html
Wenerstrom B, Kantardzic M, Arabmakki E, Hindi M (2012) Multi-tweet summarization for flu outbreak detection. In: AAAI Fall Symposium Series
Yates A, Goharian N (2013) ADR trace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In: Proceedings of the 35th European conference on information retrieval (ECIR 2013)
Yates A, Goharian N, Frieder O (2014) Relevance-ranked domain-specific synonym discovery. In: Proceedings of the 36th European conference on information retrieval, ECIR
Zaki M (2000) Scalable algorithms for association mining. Knowl Data Eng IEEE Trans 12(3):372–390
Zhu Y, Goharian N (2013) To follow or not to follow: a feature evaluation. In: Proceedings of the 22nd international conference on world wide web (WWW'13)