ORIGINAL ARTICLE

A twitter recruitment intelligent system: association rule mining for smoking cessation

Ahmed Abdeen Hamed · Xindong Wu · Alan Rubin

Received: 22 December 2013 / Revised: 11 June 2014 / Accepted: 22 June 2014 / Published online: 12 August 2014

Abstract Digital recruitment is increasingly becoming a popular avenue for identifying human subjects for various studies. The process starts with an online ad that describes the task and explains expectations. As social media has exploded in popularity, efforts are being made to use social media advertisement for various recruitment purposes. There are, however, many unanswered questions about how best to do that. In this paper, we present an innovative Twitter recruitment system for a smoking cessation nicotine patch study. The goals of the paper are to: (1) present the approach we have taken to solve the problem of digital recruitment; (2) provide the system specification and design of a rule-based system; (3) present the algorithms and data mining approaches (classification and association analysis) using Twitter data; and (4) present the promising outcome of the initial version of the system and summarize the results. This is the first effort to introduce a practical solution for digital recruitment campaigns that is large-scale, inexpensive, efficient and reaches out to individuals in near real-time as their needs are expressed. A continuous update on how our system is performing, in real-time, can be viewed at https://twitter.com/TobaccoQuit.

A. A. Hamed (✉)
Vermont EPSCoR, University of Vermont, Burlington, Vermont, USA
e-mail: ahamed@uvm.edu

X. Wu
Department of Computer Science, University of Vermont, Burlington, Vermont, USA
e-mail: xwu@uvm.edu

A. Rubin
College of Medicine, University of Vermont, Burlington, Vermont, USA
e-mail: alan.rubin@med.uvm.edu

1 Twitter data background

In the past few years, the Twitter data analysis has occupied a very large portion of research. The Twitter data has been used in a large spectrum of domains some of which are for purely theoretical research others for specific applications. This section presents some of the latest work using Twitter data.

Chang et al. (2013) utilized the TineyURL's embedded in Tweets to improve recency ranking. Huang et al. studied micro-meme phenomenon on Twitter and discussed the importance of conversational tagging practice for the larger real-time search context (Huang et al. 2010). Ghiassi et al. (2013) introduced an approach to supervised feature reduction using n-grams and statistical analysis to develop a Twitter-specific sentiments around brand-related tweets. Kim and Park (2012) investigated the role of Twitter in political conversations and debates by analyzing the ways in which South Korean politicians use Twitter. Their study examined the emergence of Twitter as user-generated communication system. Yang et al. (2012), presented a framework for analyzing and summarizing Twitter feeds. Additionally, Meng et al. (2012), presented an entity-centric, topic-oriented opinion summarization in Twitter to solve the same summarization problem. Yang et al. (2012) also analyzed tweet posts in order to make a decision about whether to retweet a post based on its interestingness. Pavlyshenko (2013) applied data mining methods to forecast events over Twitter. Ravikumar et al. (2013) analyzed tweet contents and users in order to come up with a ranking mechanism. Zhang et al. (2013) presented a novel news ranking algorithm that utilizes tweets in order to come up with news article ranking.

Eleta (2012) presented an exploratory method of studying the language choices of bilingual and multilingual tweets in relation to types of information shared. Black et al. (2012), examined Twitter as a socio-technical transport protocol. Authors presented an analysis of how researchers may approach studies of Twitter interactions. Chen (2011) studied the strength of Twitter users by analyzing the frequencies of tweeting, messaging and replying to one another. Gayo-Avello (2013) presented a meta-analysis approach to electoral prediction using Twitter data. Stringhini et al. (2013) analyzed the reputation of Twitter accounts using the count of followers. They studied the Twitter follower markets to identify the common properties of customers belong to each market. Cheong and Lee (2011) presented their terrorism knowledge-base which was constructed from civilian sentiment and response to terrorism events on Twitter.

Though, it is clear that there is much research that have been conducted using Twitter data, there is much more that we have not listed here. This brief background is meant to serve as an evidence that Twitter data is very popular in virtually any research. More specifically, in designing rule-based and knowledge-based system. Section 2 presents the motivation and background for digital recruitment and online intervention.

2 Introduction

Digital recruitment is a popular online method that has been widely used for attracting individuals who are seeking products and services. Various online services have existed over the years to assist either individuals looking for jobs, or organizations seeking individuals who can offer services of interest. In recent years, digital recruitment is transforming from passive websites such as Monster (1999), and CareerBuilder (1995) to an active form as it is the case with LinkedIn (2003). Employers have direct access to individuals' profiles and can recruit those who meet their needs. More interestingly, with Twitter being built-in to LinkedIn, employers now tweet job-postings that they will be visible by hundreds of thousands people who can apply with their LinkedIn profiles in a matter of seconds.

Recruitment in the biomedical domain is still using more traditional methods of recruitment such as phone calls, emails and advertisements on Craigslist (1995). There is room for biomedical research to use some of the online resources used by product vendors. This paper is based on a specific project to improve subject recruitment for a smoking cessation study. Participants in the study will test two ways to use nicotine patches to help them quit. For recruitment we need to reach smokers who are interested in quitting, and then prompt them to seek additional information about the study. For similar studies in the past we have used newspapers, flyers, ads on buses, and Craigslist. We also hired a vendor to manage a GoogleAdWords campaign (1995). It is inexpensive to recruit with Craigslist, but its main disadvantage is that the results may have limited generalizability, because it represents a special subsection of society—people searching Craigslist—and may not reflect society as a whole. Newspaper ads are of decreasing effectiveness with increasing costs every year. Google search ads managed by a vendor are effective for targeting audiences, but costs are high ($300–500/recruit). Flyers and bus ads are effective and of moderate to high cost ($200–300/recruit).

With shrinking research funds available, we need to find an effective and inexpensive way to recruit study participants. An additional issue is that all recruitment efforts for human research must be approved by an oversight committee [an investigational review board (IRB)]. The use of social media for recruitment is new for board members, so they may be reluctant to approve new techniques to recruit with social media techniques. So far, we have obtained permission to experiment with Twitter to identify smokers interested in quitting. During this testing, we sent public service messages on how to quit smoking instead of recruitment messages for the study. It is a challenge to create a program that will not be perceived by the board as coercive, offensive, or misleading.

Previous results As a yet-to-be published recent study, we did a combination of online and traditional recruitment and we received 2,871 responses to the recruitment efforts; 1,131 people were screened (39 %) and 249 individuals (22 %) were eligible. Of those eligible, 238 individuals (96 %) participated in the study. We anticipate that the current study will be sufficiently similar to the previous study in that we will be able to make meaningful comparisons regarding contacts, screenings, and recruitment.

2.1 Recruitment strategy and specification

Effective recruitment requires deeper understanding of the factors influencing a user response to an ad (Gupta et al. 2012). Generally speaking, for a given recruitment campaign this involves: (a) identifying interested users in the products or services; (b) catching the user in the time of need as they express their needs; (c) learning the user's take on the given recruitment message and incorporating this feedback; (d) increasing user's awareness of the products and services that fulfill their need; and (e) identifying large communities to target as a whole.

Using our innovative Twitter-based system, it is fairly easy to satisfy the criteria above and identify users who are expressing interest in quitting smoking. Twitter is a social media that supports the existence of smaller communities, which we can discover computationally and target as a whole. Using the Twitter streaming application programmable interfaces (API), we can identify those individuals who seek to quit in near real-time. Learning the feedback from users about a tweet is expressed in different forms on Twitter: (a) users can choose to follow an account or directly respond to a tweet positively or negatively using the reply feature or the messaging feature; (b) users can also favorite a tweet they view as it shows on their timeline; and (c) users can retweet a tweet and share it with their followers. This is one of the most powerful features for not only expressing how much they like the Twitter campaign, but also sharing the information with their circle of followers. Unlike Craigslist and other traditional recruitment platforms, Twitter users enjoy a sense of community. They can share ideas and trade experiences which could be very helpful in making a smoking cessation campaign successful. These expressions of ideas and experiences can flow over the network in a matter of seconds. Since our system is intended to interact with people, it is important that it does so in a more personal and human-like fashion.

2.2 Contributions

The contributions of this paper are as follows:

- We have designed a near real-time smoking cessation recruitment system using Twitter to immediately fill-in users needs. To the best of our knowledge, this is the first study that investigates smoking cessation recruitment using social media (e.g., Twitter).
- We present a non-textual, features-based classification approach that predicts whether a tweet is prestigious on Twitter's built-in features (Lists, Retweet, Verified account, Number of followers, etc.).
- We have incorporated domain expertise and engineered a rule-based system that decides which event is performed given an incoming tweet. This ensures that the system performs more intelligently in a human-like manner.
- We have demonstrated how we can derive rules from tweets that can make the system more intelligent using data mining approaches (classification and association rules).

3 Problem formulation and computational approaches

The focus of our study is to use Twitter to identify potential candidates, send recruitment messages, and increase the awareness of smoking risks and compare our approach to other online traditional methods (e.g., flyers and newspaper ads). Through an online campaign, we aim to identify Twitter users who are explicitly seeking to quit tobacco. In the following section, we formulate these tasks computationally. We use the following Twitter means to launch a campaign:

- *Explicit Tweet contents* We have inspected a large number of Tweets and manually selected very specific content to search for when a tweet is streamed.
- *User PageRank within Twitter graph* Users who have a higher PageRank are likely to be influential ones.
- *Tweet classification* Tweets are classified as prestigious or non-prestigious using classification algorithms
- *Twitter built-in gears* Retweets, Lists, and Trends are powerful means to discover communities of users. They present effective indications for how a given campaign is successful.

3.1 Proposed approach

In recent years, rule-based systems have gained much popularity due to the natural decision method that humans often follow to solve a problem (Tsakonas et al. 2004). In the medical field, rule-based systems have been widely used as medical diagnostic systems which comprise pre-engineered observations and symptoms (Wang et al. 2007; Fung et al. 1989; Aiello et al. 1995; Lhotska et al. 2001). While Haung et al. introduced a rule-based formalization of eligibility criteria for clinical trials using means of Prolog logic programming, Uzuner et al. (2008) designed a rule-based systems to identify the smoking status of patients. This research have provided us with insights to design a Twitter rule-based system that can be used to identify the eligibility of online smokers who can potentially participate in a nicotine patch study. Additionally, the domain experts who are currently using this system, have much expertise in recruitment and tobacco smoking intervention which must be used to guide the system (Hughes et al. 2013; Lee et al. 2014; Hughes 2013). Twitter, on the other hand, also imposes rules and regulations on how often a tweet can be sent, and who may or may not be contacted. In order to successfully recruit subjects, the system must adhere to the protocol of the study and knowledge engineer these rules and regulations. These facts and logistics have directly lent itself to the design of a rule-based system where rules from the various sources must be carefully observed and intelligently executed according to the appropriate scenario.

The Twitter rule-based system we present is interfaced with a recruitment algorithm that actually performs the appropriate action in correspondence with the rule fired by the system and is discussed in Sect. 5. The system operates on streaming tweets that are intercepted and processed in near real-time to identify candidate users. The system is expected to empower a Twitter account called

(TobaccoQuit) and automate most of its activities and interactions. Therefore, the system can post a tweet, mention a user in a tweet, retweet a tweet, and follow a user, also send direct messages (DM) to its followers. These various activities are explained in detail in 3.2. For the system to have a smoking cessation context, it must be seeded with smoking cessation-specific keywords. To provide unbiased input keywords to the system, we used the terms available at medical subject heading (MeSH) taxonomy, by the National Library of Medicine (NLM). MeSH descriptors can be acquired by searching the NLM MeSH Browser available online (2013). We selected descriptors from the Smoking, Nicotine, and Addiction branches. The purpose of these terms is to filter in all tweet that are relevant to our nicotine patch study. Further filtering becomes necessary because the tweets that contain the mere search keywords are noisy. A tweet such as ("I gotta quit smoking weed") contains the word "smoking" yet the tweet is not relevant. The term "weed" in this tweet makes it only relevant to marijuana smoking not tobacco smoking. Such negative words are not always apparent. It must be continuously learned by means of association analysis as described in Sect. 6.

3.2 Twitter logistics

Twitter has features, rules and regulations to ensure a good experience for its users. These rules must be observed for the algorithm to work correctly. We discuss the various features that each user can perform, and how the algorithm uses it in each computational step.

- *Tweet* This is the feature that makes Twitter the prominent social media platform it is today. A microblog of 140 characters allows the user to voice their opinion to the Twittersphere. When a user sends a tweet out, the tweet is displayed on the user's timeline and it is viewable by all of his/her followers. The tweet may also be viewed by any other Twitter user searching for keywords that match some content in the tweet.
- *Reply* When a user tweets an update, followers can directly respond to that tweet. For the reply to reach the receiver, the original user's screen-name is appended to the tweet and proceeded by the at (i.e., @) sign (e.g., @MyTwitterScreenname).
- *Mention* Similar to the Reply feature, any Twitter user can share any content with a specific Twitter user by simply mentioning their screen-name anywhere in the tweet. This does not have to be in response to a tweet.
- *Retweet* This feature is one of the most innovative features a social media platform has ever invented. When a user receives a tweet, they have a way to share it with their entire network of followers. Some tweets get retweeted thousands of times or perhaps more.
- *Lists* Twitter enables its users to group other users who share a similar interest. This mechanism is capable of creating community of users.
- *Hashtags* Twitter treats any keyword(s) preceded by the (#) as a special string. Hashtags are known for annotating tweets and making it easy to search, track and follow. Such special keywords are called hashtags.
- *Trends* Twitter shows hashtags, words, names of people, or any topic that is trending in real-time. The trending list is always updated as new events start to trend and older ones die out.

Additionally, Twitter has certain rules and restrictions that must be observed by its users. As mentioned previously, this is crucial for any algorithm that attempts to automate sending tweets, retweets or perform any other actions using the Twitter API. The following are some of the most important ones that our recruiting algorithm observes.

- *Max daily limit* Twitter restricts the number of tweets that a user can send from one account to 1,000/day.
- *Duplicate status* Twitter does not allow tweet duplicates within a given duration of time. Each tweet must be unique in content for Twitter to allow it to be posted.

4 System architecture

This section describes the architecture of our system and illustrates how online recruitment can be done in a time-aware fashion. There are three different steps: *Twitter monitor* is a software component that keeps track of the tweet streams, Lists, and Trending events and words. *Tweet analyzer* is a filtering component that queries the real-time tweets for specific phrases that explicitly indicate calls for help to quit smoking. *Event processor* is the *online* transactional component that sends recruitment messages to Twitter users among others. This component is implemented as an expert system shell. Next is an elaboration on each component individually.

4.1 System input and output

Here we describe the various system input and the expected output: (1) a set of search keyword list (e.g., smoking, tobacco, nicotine, quit). A complete list can be found in the medical subject heading (MeSH) corresponding branch. (2) A set of negative and noise keywords (weed, pot, grass, hemp, cannabis, stoner, sex, money, etc). If a tweet happens to contain any of these keywords, it is ignored. (3) Explicit sentiment that expresses the wishes to quit smoking tobacco (must quit, gotta quit, give up, stop, should). (4) A database of tweets to announce the study, promote good health and giving up are generated on the fly and annotated by trending and timely hashtags.

4.2 Twitter monitor

The system we designed is real-time recruitment software that reaches out to those Twitter users who are soliciting advice, help, or products to give up smoking. Therefore, the Twitter monitor is designed to read the streaming tweets in real-time. Since the system is concerned with smoking cessation, we designed the monitors to track those tweets that have related keywords (smoking, tobacco, quitting, addiction, cigarettes, etc). The system also monitors real-time trends (Yang et al. 2012) to keep track of emerging events and news that are encoded in English and occurring within the US. Using the publicly available Twitter REST web services (Twitter.com 2006), and Twitter4J Java wrapper API's (Yamamoto 2007), we developed the Twitter monitor component.

4.3 Tweet analyzer

Due to the massive amount of tweets received by the monitor, further analysis must be performed to filter out the irrelevant tweets. Once captured, the analyzer groups the tweets into three groups:

1. *Platinum Tweets* those tweets contain contents that solicit explicit help to quit (e.g., "I must quit smoking tobacco now").
2. *Golden Tweets* tweets that contain contents that indirectly solicit help to quit (e.g., "smoking makes me cough my lungs out").
3. *Info Tweets* all other tweets that contain useful information which can be shared with followers.

The Tweet analyzer is a simplified version of the *Text classifier* software and is based on a regular-expression dictionary look technique. We developed a home-grown component using Java Regular Expression and String manipulation. Upon the completion of this step, the tweet is passed to the event processor component along with a label (e.g., *token*).

4.4 Resources database

As mentioned above, the proposed recruitment system operates mainly on streaming tweets. Nevertheless, it needs a backend database to be fully automated. Our team has created a large number of pre-prepared tweets to interact with the target Twitter users. One type of this interaction is to perform soft-recruiting by periodically announcing our recruitment services to followers and the outside world. These tweets contain the following contents: (1) uplifting informal messages to share; (2) a 24/7 voice service number for people to call; (3) a tiny URL that links users to the website of the services; and (4) several hashtags that we selected from existing hashtags as well as some of our own. The hashtags are used based on the context, day of the week, and ongoing events. (e.g., #Lent, #TGIF #FollowFriday to use on Friday, and #HappyMonday to use on Monday) are appended to the tweet body dynamically. In addition, there is a new type we call direct recruitment for the tweets that are sent to a Twitter user in response to a tweet that seeks help. Another type of tweet is one that enables democratization of knowledge and shares the most important news, research studies and other services to increase the awareness of tobacco smoking risks. All three types of tweets are generated on the fly using the links, pre-prepared tweets and hashtags. Table 1 shows some sample recruitment messages we send in different contexts.

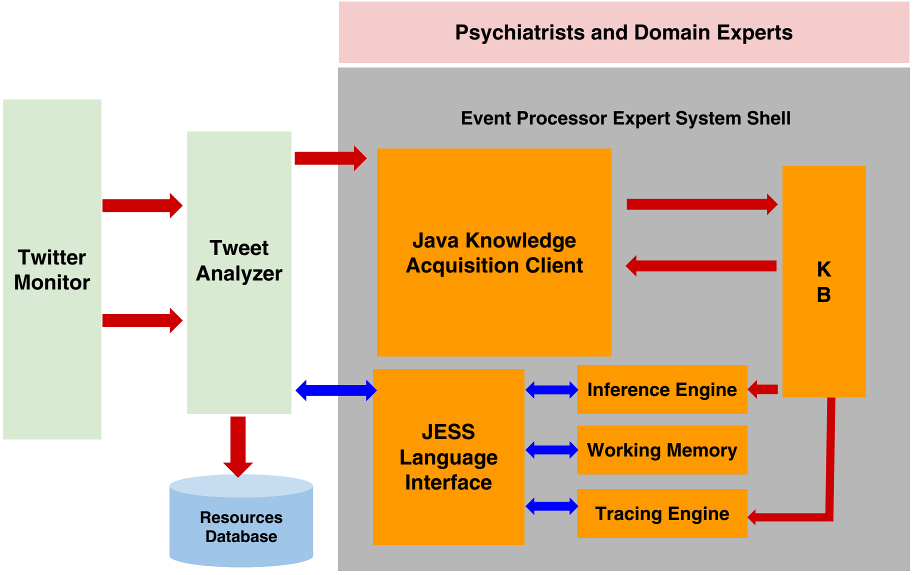
Table 1 Pre-prepared recruitment sample Tweets

Tweet type	Tweet text
User Tweeting	I want to stop smoking again.
User Tweeting	Why is it so hard to quit smoking cigarettes :(Someone help me ...: 17 hours ago ... I have good reasons to quit.
User Tweeting	I should give up smoking cigs for lent .... SIKE !
Soft recruitment	Make friends with your lungs; quit smoking!
Soft recruitment	Step out of the past and into a smoke-free future.
Soft recruitment	Breathe easier. Quit smoking!
Direct recruitment	You can quit now. Be smoke free quickly. call #877-437-6055 #smokefree
Direct recruitment	!!!!!Good idea! It's time to quit. Here's help: bit.ly/11Z2GAy #lent
Direct recruitment	Take action to help a loved one quit: bit.ly/11Z2GAy#socialcare

4.5 Event processor expert system shell

When a tweet is classified to be Platinum, Gold or Info, a token is sent to the event processor component to perform an action. Based on what the token is, an action is fired. This component is powered by an algorithm that performs a sequence of tasks until the event is processed. Starting with the token received, the algorithm activates an action and dynamically binds the action with database contents, checks the current rate for this particular type of event. It is common to mine patterns in data streams based on streaming textual

Fig. 1 Twitter recruitment system architecture



feature selection methods (Yu et al. 2012). We follow a similar approach here but we look for non-textual features (e.g., user number of followers, number of friends, on a specific Twitter list). Based on the existing features associated with each tweet, the tweet gets a final score. Additionally, the processor crawls any tiny URLs that exist in the received tweet body to compute a local approximation of PageRank to incorporate into the prestige of the tweet. For this task, we have utilized an open-source web crawler Java library called Crawler4J (Ganjisaffar 2012). Figure 1 shows the architecture and the various components. The algorithm is explained in the upcoming sections.

4.6 Knowledge base

At the heart of the recruitment system lies a rule-based component that communicates with the textual phrases in the tweet body. When a tweet is analyzed by the Tweet analyzer component, the output is sent to the rule-based system to decide which task should be performed. When this is completed, a decision is made and a token is sent back to the event processor. This system comprises five main rules that correspond directly to the Twitter features that the recruitment system is based on. We implemented this system using JESS, a Java expert system shell (Hill 2003).

4.6.1 *Tweet rule*

This is a direct implementation of the Tweet feature to soft-recruit individuals without being "spammy". Since it is able to read and analyze the streaming tweet content in real time, the tweet rule is able to decide when to tweet a message to the world if someone is seeking help quitting tobacco. Although the tweets sent out are not a direct reply to the user seeking quitting information, it contains Twitter hashtags that will match the sender's hashtags. This will increase the chances of finding our tweets while this particular user is still active on Twitter.

Algorithm 1 Tweet Rule Implementation.

```
(defrule action-soft-recruit
(twitter-user (language "en")
  (screen-name ?screenName))
(raw-tweet-info
  (id ?tweetID)
  (text ?tweetText))
(test (and (not
     (contains-retweet-keywords
       ?tweetText))
  (not
     (contains-article ?tweetText))))
=>
(assert (recruitment-action
  (action "soft-recruit")
  (tweetID ?tweetID))))
```

4.6.2 *Reply rule*

Currently, the reply feature looks at simple specific tweets that contain messages from users who are explicitly soliciting help to quit tobacco. Once identified, the rule is activated and the action is triggered as "Reply". All facts in the working-memory of the knowledge-based system are bound to this Reply only. This makes composing the reply to the tweet possible. The rule makes use of another local rule which decides if a tweet has the basic and specific contents. The reply rule must match this pattern to perform basic filtering before the action is activated.

Algorithm 2 Reply Rule Implementation.

```
(defrule reply-action
(golden-tweet-info
  (id ?tweetID)
  (text ?tweetText))
(twitter-user
  (language "en")
  (screen-name ?screenName))
=>
(assert
  (recruitment-action
     (action "mention")
     (tweetID ?tweetID))))
```

4.6.3 Retweet rule

Twitter's Retweet feature (aka RT) is one of the most influential features of Twitter for getting the word-of-mouth circulated quickly to so many users. RT is a very powerful tool such that when a user searches for a tweet or receives one, they can choose to share it with their followers via this feature. Once a tweet is retweeted it becomes visible to all the original sender's followers. After identifying a relevant tweet, this rule further decides whether it should be shared in the form of RT or not.

Algorithm 3 RT Rule Implementation.

```
(defrule action-retweet
(twitter-user
  (language "en")
  (screen-name ?screenName))
(raw-tweet-info (id ?tweetID)
  (text ?tweetText))
(test (and (contains-retweet-keywords
  ?tweetText)
(contains-article ?tweetText)))
=>
(assert
  (recruitment-action
     (action "retweet")
     (tweetID ?tweetID) )))
```

5 Twitter recruitment algorithm

The purpose of this algorithm is to exhibit intelligent behavior that is considered acceptable by the general public (Twitter users). The algorithm should not propagate spam, follow or recruit irrelevant users. It also must obey Twitter rules in order for the Twitter account to remain lively and influential. More importantly, it must demonstrate courteous behavior to its followers and to the users who are seeking help to give up smoking. For example, if a recruitment message is sent immediately after a tweet that is identified to be relevant, it elicits a negative reaction from users. Therefore, a delay function is used to relax the behavior of our system when it tweets from the @TobaccoQuit Twitter account. Another important goal of the algorithm is to enable the sharing of knowledge by sharing tweets that may contain important information on the risks of lung cancer, new research studies on smoking, and uplifting experiences shared by former smokers.

When a streaming tweet is intercepted, it is immediately checked for its language(must be English) and searched for the basic set of keywords. If the algorithm recognizes the tweet as relevant, The algorithms communicates such information and other features with the rule-based system components to make a decision. The rule-based system gets populated with rules stored in its knowledge base. When a decision is made, a token is sent back to the algorithm (RT, Mention, Tweet, etc). It is up to the algorithm at this time to actually perform the action if Twitter rules permit. For instance, if the system reached the quota of sending a favorite action algorithm will not perform the action. It performs a Retweet (RT), a Mention, or a simple Tweet action. Algorithm 4 shows the pseudocode.

Algorithm 4 Data Stream Recruiting Algorithm

```
Input:
$W$: basic keyword list
$A$: action {Direct, RT, Soft}
$r$: prestige threshold
Description:

Foreach tweet $t$
  search(t, W)
  If relevant
    score = prestige($t$)
    If action(RT) AND score $>$ $r$ and
      perform_action(RT)
    Else If action(Direct)
      perform_action(Direct)
    Else
      perform_action(Soft)
    End If
  End If
End Foreach
```

Beside the basic set of keywords to filter out the relevant tweets, the algorithm expects action tokens from the Rule based system: (1) *Direct* token is an string that signals a direct recruitment action originally made by the rule-based system; (2) *RT* token to signal an action that retweets the current tweet and shares it with followers of the system's main Twitter account; and (3) *Soft* token signals a soft recruitment action. The algorithm also accepts a prestige threshold which is set by the domain experts to determine the prestige score of the tweet sender. This score is necessary for the algorithm to decide whether to fire a Retweet rule.

Based on the rules fired by the JESS engine (RT, Direct, or Soft) the algorithm decides whether to actually fire the rule. If a rule is fired, the algorithm is responsible for making the execution occurs. The body of the algorithm is essentially an infinite loop that analyzes tweets, one at a time. The *search* function, searches the tweet text against the list of keywords to determine its relevance. If a tweet is found relevant, it is checked against a classification algorithm for prestige. Depending on the result of classification, the tweet is retweeted. This results in sharing the tweet with the follower of out account to share news articles or medical advice. If the action made by the engine is ("Direct"), then a direct recruitment strategy is executed and a *Mention* is sent to the sent to this particular user to let him/her know about the services provided and the nicotine patch study. If the tweet is relevant and none of first two conditions is satisfied, then the system sent a ("Soft") recruitment. This can be simply be either marking the incoming tweet with *Favorite* or sending a tweet from the database of prepared tweets.

5.1 Retweeting Tweets by classification algorithms

Various studies have attempted to measure the prestige of a tweet. Some studies assumed that if a tweet has a URL then it is important. Although this might be true to some extent, further analysis is necessary. Other studies considered the number Retweets to a tweet as signifying the importance of a tweet. This is indeed interesting but an agent program can simulate this behavior and circulate spam using RT. However, the above methods have inspired Yang et al. (2012) and Yamaguchi et al. (2010) to recycle the idea of using the RT feature as an indication of a high prestige.

The act of sharing a tweet with followers is called retweeting (RT). People RT other peoples' tweets for various reasons: (a) a tweet may contain a link that points to important article or a page on the web; (b) adopting or supporting an opinion expressed in the original tweet; (c) the tweet is sent by an influential person (e.g., Michelle Obama); and (d) other reasons. These characteristics can be represented in terms of non-textual features and can be used to classify a tweet as RT candidate using classification algorithms.

- *Number of followers* The more followers the sender of the tweet has the more influential and likely to be retweeted (Stringhini et al. (2013)). This feature can be quantified using the PageRank algorithm. A theoretical background is presented below.
- *Verified accounts* Twitter identifies certain accounts as verified. Such accounts belong to people who are known in business (Richard Branson), politics (Muhammad Morsi), celebrities (Celine Dion) among others
- *Web links* Tweets may contain a link, also known as a tinyURL, which is a pointer to a physical web resource
- *Trusted users list* People or experts in a particular domain and known personally to the person who tweets. A person may trust his own medical doctor, professors, academic advisors. Twitter supports such idea by providing List device for its users to trust, mark for spam, and mark for favorite. This is a feature our automatic approach maintains by looking up a Trusted list for the RT feature.

5.1.1 *Justification of number of followers feature*

Measuring the user's PageRank is a good indication of how prestigious the user is. However, PageRank requires the ranks of in-degree users (the followers) and out-degree (friends). Twitter imposes limits on not allow querying each tweet for its user and retrieving the handles of its followers. Including the number of followers (in-degree) has been already proven to be equivalent to PageRank. The following is the theoretical background to show that the in-degree is a good approximation of the global PageRank.

PageRank score This score is determined by summing up the PageRank scores of all pages that point to i (Wu et al. 2007; Wu and Kumar 2009). The score of page i [denoted by $P(i)$] is defined by:

$$P(i) = \sum_{(j,i)\in E} \frac{P(j)}{O_j}$$

s.t. O_j is the number of out-links of page j.

In-degree PageRank score The average PageRank of a page with in-degree k_{in} can be well approximated (Fortunato et al. 2005, 2007) by the following closed formula:

$$p(k) = \frac{q}{N} + \frac{1-q}{N}\frac{k_{in}}{\langle k_{in}\rangle}$$

s.t. N is the total number of pages, $1-q$ is the so-called damping factor, and q is the probability of jumping from a node to another in a random walk.

Table 2 Tweet classification features for automatics retweet

Handle	Follower no.	Trusted users	Verified	URL	Prestigious?
JaniceKronebusc	0	No	False	Tiny	NRT
GeoffreytheFish	0	No	False	Notiny	NRT
RiciclAr	1	No	False	Tiny	NRT
rickyrascalj86	1	No	False	Notiny	NRT
bigsh125	1	No	False	Tiny	NRT
hurtandhelpless	7	No	False	Notiny	NRT
TopMaster434	8	No	False	Tiny	NRT
MilnAlbert	11	No	False	Notiny	NRT
prettystuddDae	18	No	False	Notiny	NRT
MadieMelecio	24	No	False	Tiny	NRT
DebiopharmNews	35	No	False	Tiny	NRT
tips2quit	23	Yes	False	Notiny	RT
teamoncology	7,939	No	False	Notiny	RT
KXAN_News	45,663	No	False	Tiny	RT
CDC_Cancer	52,000	No	True	Tiny	RT
MakeAWish	162,960	No	True	Tiny	RT
Telegraph	497,900	No	True	Tiny	RT

We treat each tweet as its own local directed graph DAG as follows: $\forall$ Twitter user u_i, $\exists\, U = \{u_i, \ldots, u_k\}$ such that U is the set of k nodes. The relationship between u and U is established by Twitter's *Following* relationship. By applying the in-degree approximation formula of the PageRank algorithm, we get the score of user prestige in the Twitter network.

5.1.2 Retweet classification data and results

We collected 5,000 tweets, each of which must contain one or more search keywords mentioned above. We eliminated the tweets that came in with wrong encodings or language which could not be read. We also eliminated the tweets that were a RT of original tweet to remove any bias in the training set. The remaining tweets were labeled as prestigious or not using RT or NRT labels accordingly. We then selected a single record for each discriminant feature to represent the positive cases. We also selected tweets that are clearly not good candidates for sharing or retweeting. This resulted in 17 instances 11 negative instances and 6 positive instances shown in Table 2.

The training data above were used by four classification algorithms [Naive Bayes, CART (Wu et al. 2007; Wu and Kumar 2009), and Random Forests (Breiman 2001) as statistical-based classifiers)], J48 is Weka's representation for the C4.5 decision tree classifier (Wu et al. 2007; Wu and Kumar 2009). Figure 2 show the performance of three different measures produced by the four different algorithms (precision, recall and F-measure) for both the NRT and RT, respectively. It is clear that Naive Bayes algorithm performed the lowest precision score (84 %) for identifying the negative instances. Random forests, on the other hand, produced a (100 %) precision score when identifying both positive and negative instances.

5.2 A comprehensive running example

Using the Twitter4J Java API, the monitor component captures tweets that has the input keywords (e.g., smoke, smoking, cigar, cigarette, tobacco, nicotine, etc). Tweets that do not contain one or more of these keywords are simply ignored. Tweets that has at least one keyword is further examined for relevance and categorization. It is trivial to understand that not every tweet that contains the word smoke is related to smoking (e.g., "the battlefield beta is LIVE! *smoke effects, fireworks, airhorns, Lil John shouting "TURN DOWN FOR WHAT" conclude the EA presentation*"). It is also trivial to see that not all tweets that contain the word smoking is related to individuals who are seeking help to quit smoking (e.g., "Smoking weed is a hobby not an addiction"). These type of tweets generate noise which must be eliminated. The Tweet Analyzer component dissects each individual tweet and group it into (Platinum, Golden and Informational). A tweet such as (e.g., "I want to quit smoking but continue doing other stuff", 'I want to quit smoking before I leave August 27th') are categorized as platinum. This is due to the presence of ("quit smoking") literals and the absence of negation. Such tweets trigger a Mention action which would result in a tweet that will be sent to the sender's timeline inviting him/

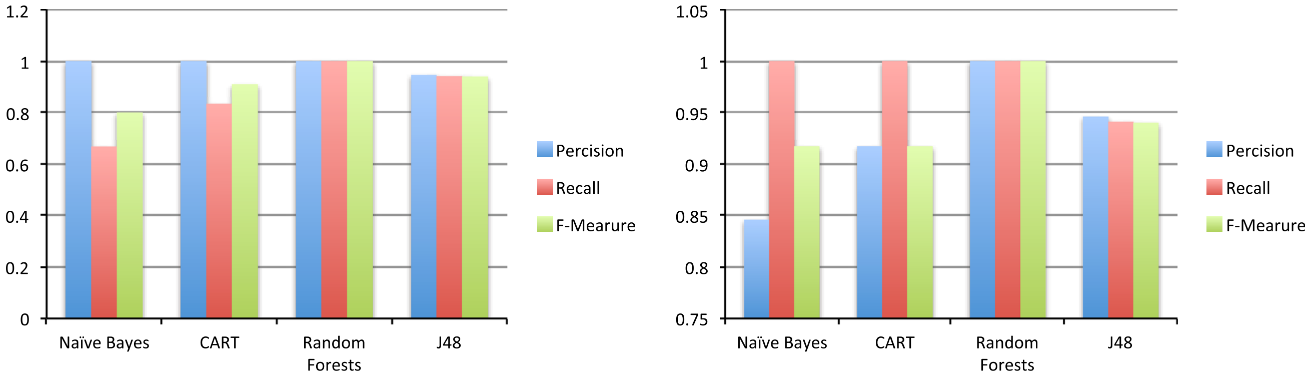


Fig. 2 Performance analysis of four classification algorithms

her to take the screening to check illegibility for receiving a nicotine patch by regular mail. The TobaccoQuit Twitter account shows various mention actions in response to platinum tweets ("@USERNAME Step out of the past and into a smoke-free future. http://bit.ly/156Irhn 877-437-6055 #ResearchStudy #SmokeFree") given that "@USERNAME" token is replaced by a real twitter handle.

The following tweet ("I'm happy I quit smoking cigarettes but damn is it haft to not want them") is classified as golden because it contains ("quit smoking" + "cigarettes") literal and also contains a negation. The negation make the tweet ambiguous, therefore this user does not get recruited directly by a personalized tweet as in the case of a (@Mention) but rather a soft recruitment using a Favorite action. This triggers the system to mark the tweet as Favorite, which is similar to the Like feature in Facebook. This action appears on the Interaction window of the user and it gets noticed. It is up to the user to decide to visit the TobaccoQuit account and extract the phone number and take the survey. Hence, it is called soft recruitment. Incoming tweets that make the criteria of information tweeted triggers a RT action and will be retweeted as described in Sect. 5.1 above.

The TobaccoQuit account also posts regular tweets based on traffic generated by people if the account did not match its quota for a given duration. Such quota information is kept track of by the monitor component. Such tweets are stored in a resources database which is populated by domain experts who generate new tweets and approve its use by the IRB. Before a tweet is sent, it is comprised on the fly from previously preprepared tweets and hashtags. Hashtags are inserted based on the day of the week (#TGIF for Friday) and holidays (#HappyHolidays). Such rules are encoded in the knowledge-base (KB) component of the rule-based systems. They are triggered in the presence of a soft recruitment action.

It is also worthwhile noting that while the system operates to performing recruitment actions, the account also receives new interaction from the outside world (e.g., Follow, Mention, Reply). If the account is ("Followed") by another user, a direct message rule is triggered which results in sending a private email to the new follower informing them about the study and sharing the contact information. Direct messages are restricted by Twitter and can only be sent to the Followers. Such rules and restrictions are encoded in the knowledge base (KB) component of the expert system shell. All actions and activities happen in near real-time. It is important to note that the system does not perform a ("Reply") action to the inquiries we @TobaccoQuit receives for various reasons: (a) The system was designed to recruit people to the study, but not to perform intervention. Answering open-ended questions can only be addressed by domain experts and the purpose of this system is to recruit not to entirely replace domain experts; (b) Twitter does not expose APIs to respond to the interactions occurring.

6 Knowledge acquisition using association rule mining

Twitter is a dynamic social media where its contents are constantly evolving. Having a fixed set of rules deliver by domain experts will not suit this ever-evolving nature. Rules must be acquired from the tweets themselves in addition to the knowledge by the experts. We performed a series of rule mining experiments to come up with the rules we add to the knowledge base. It is essential to note that tweet contain various types of contents some of which is plan text, URL links, and special type of words are called hashtags. A hashtag is a word or a collections of words proceeded by the symbol (#) and concatenated together with underscores or using a upper-lower case convention. Hashtags carry various semantics and which can immediately identify a tweet to be relevant or irrelevant. For instance, a tweet such as ("#420 is over but im still high :), I better quit smoking soon!") is irrelevant to our nicotine patch study, though a bigger chunk of this tweet suggests that this person may be seeking help to quit smoking.

Hashtags such as (#420) are seasonal and related to temporal events. (The April, 20th event observed by marijuana smokers in North America). Tweets that contain this particular hashtag will only appear around the month of April. A dataset gathered in January or February may not contain instances of this hashtag. Nevertheless, the incoming traffic of such an event must be considered by our monitors. The system is intended to run 24 h and 7 days all year around. Using means of classification and clustering is limited in this case and can not accommodate this volatile nature of hashtags. We believe that approaching this problem is best approached means of Association Analysis and rule mining in general as opposed traditional means of machine learning. Additionally, association rules can also be extended and represent as networks. This particular notion has been exploited by Hamed and Wu (2014) in constructing keyword-hashtag networks (K-H) networks, which can be further mined using pattering mining algorithms on graphs (Yan and Han 2002).

The next sections explain how to we use experiment with hashtags and how to evaluate valuable rules using Apriori, and Association Rule Mining algorithm. These experiments are the background foundation for the network construction and mining explained (Hamed and Wu 2014). It is worthwhile mentioning that various datasets of various domains in addition to the smoking cessation were used to test how more complex rules can be mined. Particularly, the used four heterogeneous Twitter dataset from the drugs, cars, sentiments and smoking domains to identify significant patterns, which could not be identified by traditional association rule mining. Additionally, we designed a path-mining algorithm to operate on the (K-H) networks called *HashnetMiner*, currently in review, to identify unknown drug interaction and marijuana–drug interactions using only tweets. These newly identified rules are stored in a database as a first step of to populate our rule-based system to perform more advanced online intervention. Figure 3 shows the basic idea of how a single hashtag (#relief) can connect drug names and suggest association or interaction. The figure to the left shows how pairs of drugs are connected using association rule mining (tums and advid) and (ibuprofen and protonix) before hashtags were integrated. The figure to the right shows a how the (#relief) hashtag connects each pair. These more complex associations are further explored using *HashnetMiner* to identify which paths are significant. This algorithm uses two different networks to compare, derive, and report the novel and unknown interactions.

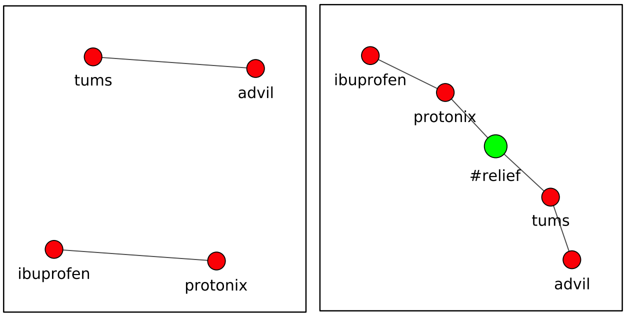


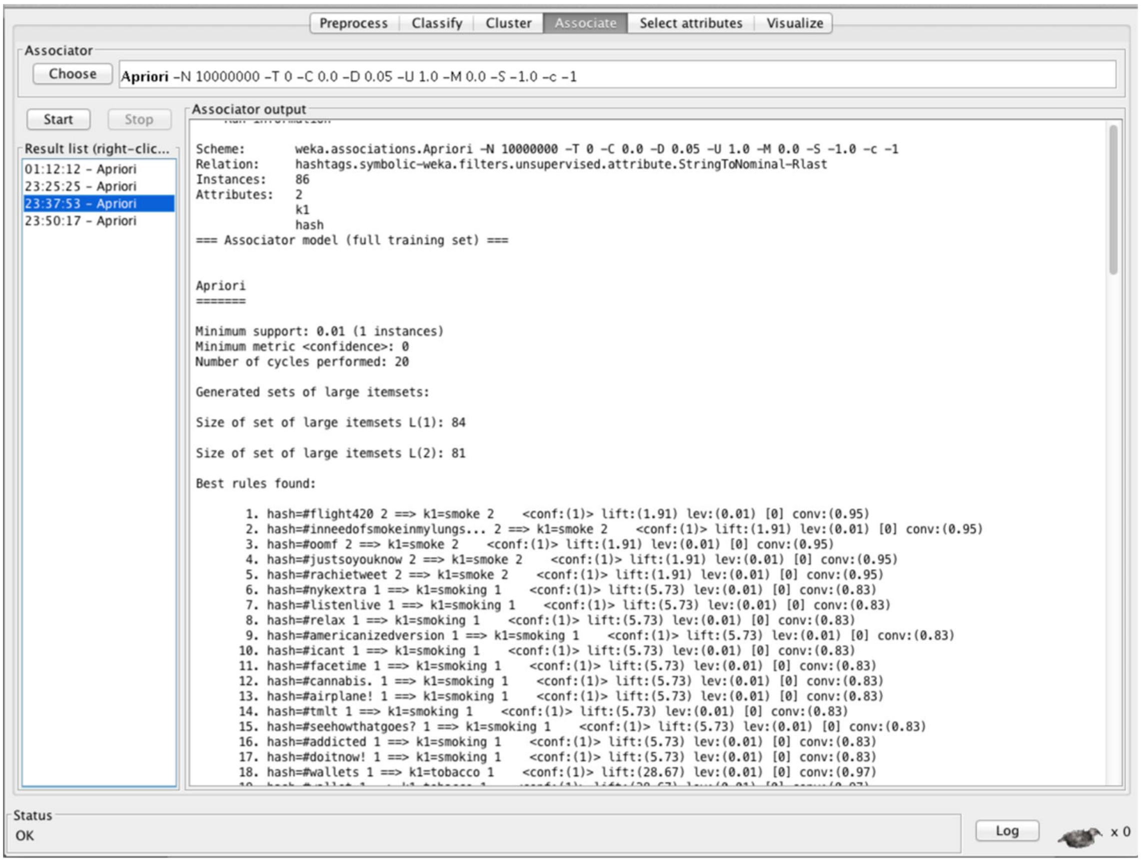
Fig. 3 The role of hashtags connecting drug name

6.1 Data description and gathering

Collected a two sets of streaming tweets captured by Twitter's streaming API. First set is gathered using ten search terms: (quit, quitting, smoking, nicotine, patches, smoke, cigarette, cig, cigs, ecig, marijuana). Second set of tweets is a superset of the first one and comprises 30 terms which include the terms from the first set. Given those set of keywords, it is possible to get tweets that contain hashtags that are identical to the search keywords (e.g., patches and #patches). It is also possible to see entirely different hashtags that may not have any syntactic similarity with the input search keywords (e.g., smoking and #weed). Our method distinguish between these two types of hashtags to see if one is better than the other and quantify that. According to our method of analysis, a tweet such as ("you can stop up smoking now, free #patches call #800quitnow") generates the following records for the given keywords we used: (1) (smoking, #patches), (2) (smoking, #800quitnow). We treat such instances as association transactions. When a large set or tweets are analyzed, they also generate a large set of transactions which we analyze using Apriori Algorithm (Agrawal and Srikant 1994). We collected two data sets from a 10,000, 25,000 tweets. Records resulted from each type is formatted using Weka's ARFF format to analyze using the Apriori algorithm. These experiments are designed to expose associations among the given keywords we are interested in and the ever-emerging hashtags that are not known in advance. Association between keywords such as (smoke, cigar), or (cigar, cigarette) are already known since we know from the MeSH taxonomy that such keywords are related. However, this very fact can also be used a ground-truth to benchmark the association resulting from (keyword, hashtag) transactions. This is explained in much more detail in Sect. 6.5 where experiments show that the association between (keyword, keyword) are not as common when compared with (keyword, hashtag). The output of some of experiments is shown in Fig. 4.

6.2 Setting up minimum support

Finding frequent item sets in tweets is a challenging task. This is due to the fact that tweets encompass a large spectrum of topics. Each topic is covered by a



Associator

Apriori -N 10000000 -T 0 -C 0.0 -D 0.05 -U 1.0 -M 0.0 -S -1.0 -c -1

Result list (right-clic...

01:12:12 - Apriori
23:25:25 - Apriori
23:37:53 - Apriori
23:50:17 - Apriori

Associator output

```
Scheme:       weka.associations.Apriori -N 10000000 -T 0 -C 0.0 -D 0.05 -U 1.0 -M 0.0 -S -1.0 -c -1
Relation:     hashtags.symbolic-weka.filters.unsupervised.attribute.StringToNominal-Rlast
Instances:    86
Attributes:   2
              k1
              hash
=== Associator model (full training set) ===


Apriori
=======

Minimum support: 0.01 (1 instances)
Minimum metric <confidence>: 0
Number of cycles performed: 20

Generated sets of large itemsets:

Size of set of large itemsets L(1): 84

Size of set of large itemsets L(2): 81

Best rules found:

      1. hash=#flight420 2 ==> k1=smoke 2    <conf:(1)> lift:(1.91) lev:(0.01) [0] conv:(0.95)
      2. hash=#inneedofsmokeinmylungs... 2 ==> k1=smoke 2    <conf:(1)> lift:(1.91) lev:(0.01) [0] conv:(0.95)
      3. hash=#oomf 2 ==> k1=smoke 2    <conf:(1)> lift:(1.91) lev:(0.01) [0] conv:(0.95)
      4. hash=#justsoyouknow 2 ==> k1=smoke 2    <conf:(1)> lift:(1.91) lev:(0.01) [0] conv:(0.95)
      5. hash=#rachietweet 2 ==> k1=smoke 2    <conf:(1)> lift:(1.91) lev:(0.01) [0] conv:(0.95)
      6. hash=#nykextra 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
      7. hash=#listenlive 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
      8. hash=#relax 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
      9. hash=#americanizedversion 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     10. hash=#icant 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     11. hash=#facetime 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     12. hash=#cannabis. 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     13. hash=#airplane! 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     14. hash=#tmlt 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     15. hash=#seehowthatgoes? 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     16. hash=#addicted 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     17. hash=#doitnow! 1 ==> k1=smoking 1    <conf:(1)> lift:(5.73) lev:(0.01) [0] conv:(0.83)
     18. hash=#wallets 1 ==> k1=tobacco 1    <conf:(1)> lift:(28.67) lev:(0.01) [0] conv:(0.97)
```

Status
OK

Fig. 4 Association rule mining using Apriori in Weka

wide range of users from different cultures, backgrounds. Clearly this is different from the fixed number items people can buy from any grocery store or online on amazon.com. We found that a very small fraction of minimum support is sufficient to expose the local trends about the topics that we are interested in (i.e., smoking cessation in this case). We have experimented with the following minimum support values (0.01, 0.01, 0.0001). This is sufficient to expose interesting patterns and associations found in the dataset analyzed. Table 3 shows a comprehensive list of parameter values. We performed three consecutive experiments on each data set by fixing the minimum-support parameter and continuously dropped the minimum confidence with a small fraction to measure the performance of each type. Table 3 shows the various settings of minimum support and confidence of the experiments.

6.3 Experimental results

Figures 5, 6 show the comparisons of confidence using difference support and confidence levels. The curves on the left show both the average confidences of each experiment. The horizontal axis reflects the a single parameter setting of an experiment and vertical reflects the average confidence of all rules learned form this experiment. The curve in blue shows the average confidences for the associations of keywords and their identical counterpart hashtags. The curve in red shows the average associations for keywords and the entirely new hashtags. The histograms in blue reflect confidence gain for data set 1 and the histograms in green reflect the confidence gain in the large data set. The previous types of experiments have demonstrated that there is indeed gain in incorporating hashtags to generate intelligent rules. However, the number of rules generated depends on the support confidence thresholds. Up to this point, the only way to identify significant rules and prune out not useful rules is manually using the domain experts. This rule selection step becomes a bottleneck and cripples the system from functioning at its best. A more automated approach is desperately needed to get rid of the manual labor. Readers of this paper are highly encouraged to read how this issue is computationally addressed using our K-H networks we briefly introduced in Hamed and Wu (2014).

Table 3 Experiments settings

Min Supp	Min Conf												
0.01	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.001
0.001	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.001
0.0001	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0.05	0.01	0.001

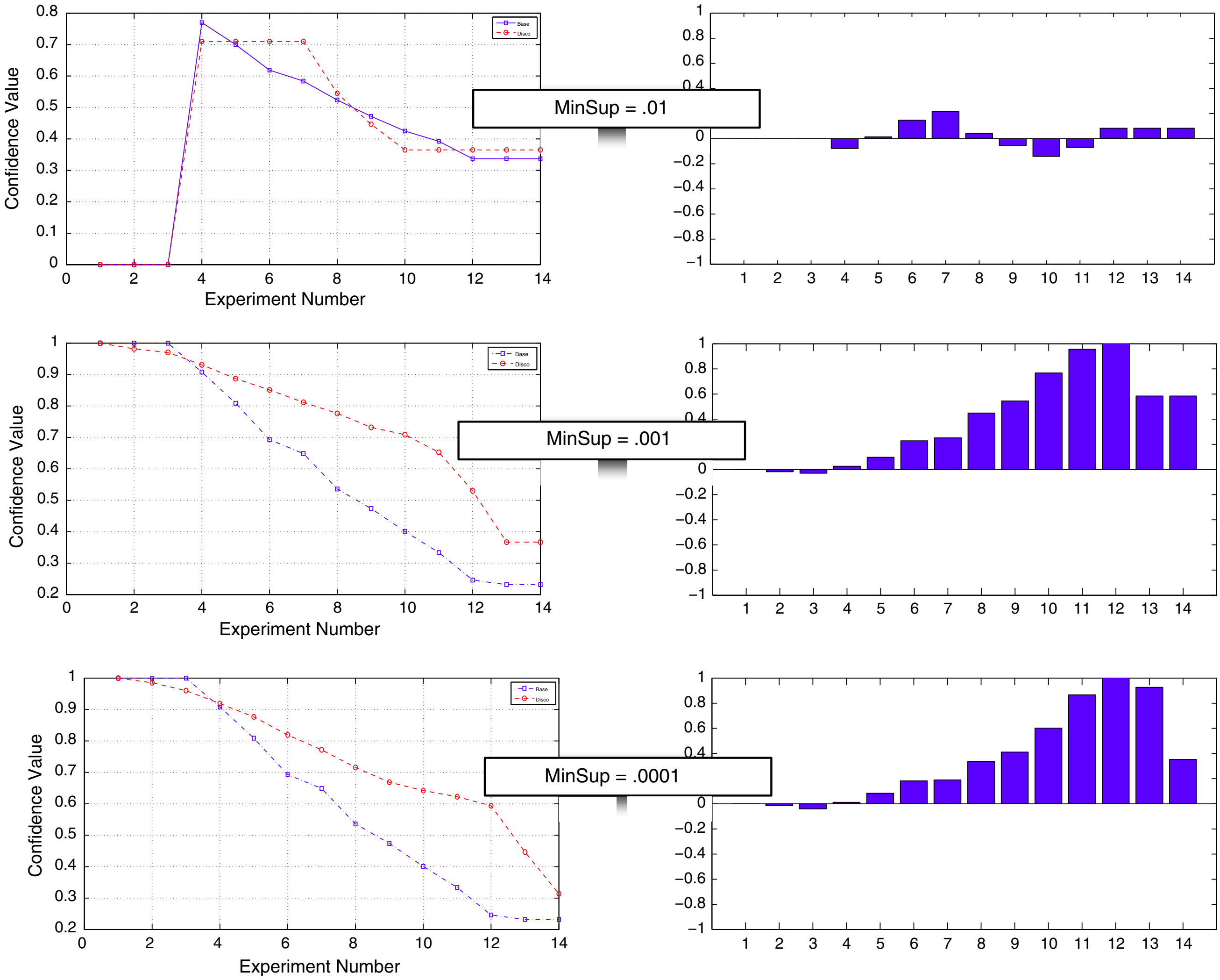


Fig. 5 KK vs. KH dataset 1 experiments with three minsup levels showing the gain ratio favoring KH

6.4 Identifying significant hashtags

The previous types of experiments have demonstrated that there is indeed gain in incorporating hashtags to generate intelligent rules. However, the number of rules generates depends on the support confidence thresholds. Up to this point, the only way to identify significant rules and prune out not useful rules is manually using the domain experts. This rule selection step biomes a bottleneck and cripples the system from functioning at its best. A more automated approach is desperately needed to get rid of the manual labor. This section introduces a ground-truth approach to overcome this difficulty and identify significant rules produced by any association analysis algorithm (e.g., Apriori). This idea establishes a standard way of comparing how much knowledge is gained (in terms of how many rules discovered) in both the absence and the presence of hashtags experimentally. We need a basic set of assumptions to perform these new experiments.

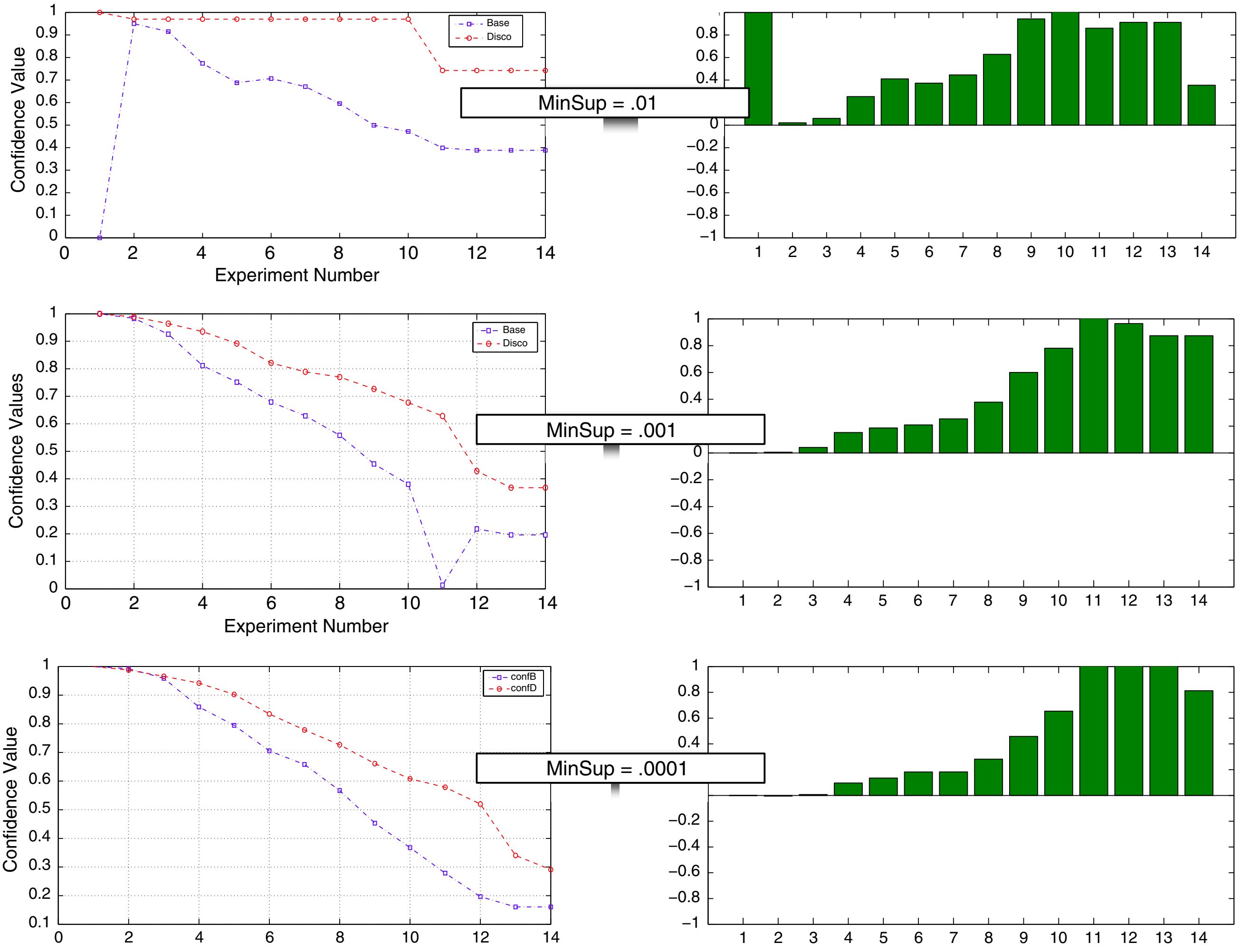


Fig. 6 KK vs. KH dataset 2 experiments with three minsup levels showing the gain ratio favoring KH

1. Tokens that make up a taxonomy concept are mutually associated. For example, the concept "Tobacco Smoking" makes the keyword Tobacco and the keyword Smoking mutually associated.
2. Two keywords are mutually associated if both co-occur in the same taxonomy and co-occur in the same tweet.
3. A hashtag and a keyword are mutually associated if both co-occur in the same tweet.
4. Experiments do not distinguish the semantics between a taxonomy keyword and its identical hashtags form (e.g., #smoking, and smoking have the same semantic). While, the experiments do distinguish the syntactic forms.

6.5 Experiment description

Our intent of this paper is to discriminate between the novel and significant rules and rules that are already known and expected. This is done by using means of identical hashtags, which are an exact match to the given input search keywords. For a given set of tweets, we mine association rules of terms that co-occur in the dataset. Considering the "Smoking" branch of the MeSh taxonomy, we can state an *essential observation*: terms comprise one or more keywords (e.g., Tobacco Smoking). There are various compound terms under the same branch: (e.g, "Cigar Smoking", "Hookah Smoking", "Pipe Smoking"). This reveals the strong association between the (Tobacco) as an individual keyword and "Smoking" as another, and the same holds for Cigar and Smoking. Analogously, we can claim that there is also an association between both (Tobacco) and (Cigar) keywords, but with less strength. It is trivial to see that the rules that might hold using any given set of tweets will be one or more of the following rules: (Cigar ↔ smoking), (Cigarette ↔ Smoking), (Hookah ↔ Smoking), (Tobacco ↔ Smoking), (Cigar ↔ Cigarette), (Cigar ↔ Hookah), (Cigar ↔ Tobacco), (Cigarette

↔ Hookah), (Cigarette ↔ Tobacco), and (Hookah ↔ Tobacco). It is also trivial to imagine that rules may contain an identical hashtag on either side. Rules such as (Cigar ↔ #Tobacco), (#Cigarette ↔ Hookah) are possible to mine if the database transactions support it. We aim to show that more rules are discovered when identical hashtags are involved as we compare with the rules mined from the keywords-alone counterpart (i.e., the ground truth baseline).

We used two small datasets to, the synthetic and real tweets, the results favored the transactions that included identical hashtags. Figures 7, 8 show the outcome of identical hashtags experiments, for both synthetics and real dataset, respectively. The horizontal axis is reserved for each possible rule and shows whether a rule was discovered or not. The vertical axis is to show the confidence level achieved for each rule if the rule was discovered. When a rule is not discovered, no corresponding bar appears. On the one hand, Apriori discovered 60 % of the expected rules from the synthetic dataset, while it discovered 26 % of the rules in the absence of hashtags. Using the real tweet dataset, Apriori discovered 40 % of the rules when hashtags were used but only 13 % when they were absent. Even through the results from real tweet dataset contained fewer rules than the synthetic tweets, they still both supported our original intuition. Additionally, the results demonstrate that hashtags are more commonly used than taxonomy terms. It is apparent that people on Twitter seem to be more interested in *hashtagging* tweets than using a plain English word. Moreover, when comparing the synthetic to the real-tweet datasets to examine rules discovered using just taxonomy terms alone, ratio of (1:2) was found. Clearly, the results show that the use of words is much less common than we originally anticipated. These basic observations make it evident to grasp the role played by hashtags on Twitter. It is evident that any scientific research that aims to study the behavior, response or reaction of a particular Twitter community must integrate hashtags to capture the full picture.

As stated previously, hashtags on Twitter do not only appear in identical forms to known domain keywords, but they also compound terms, acronyms and symbols (e.g., #TGIF, #YOLO). The first hashtags is a very well know shorthand Twitter users embed in their tweets on Fridays (ThankGod it is Friday). The other tweet is a short form of (You Only Live Once) idiom. Comparing rules of ground-truth against rules (with keywords and identical hashtags) against rules that include all other hashtags is an effective way of identifying novel rules after the generation of the rules. This basic idea inspired us to form the (K-H) networks and the design the *HashnetMiner* algorithm.

7 System performance and evaluation

The system is currently running on a test environment at @TobaccoQuit. This section gives an account of how the system is performing. We have tweeted about 23K tweets, some of which are our own messages from the database, replies and mentions. We have about 1,100+ followers, while the average Twitter account has 126 based on the Twitter statistics of 2012 (Pring 2012). In a period of 2 months, the account has accumulated more followers than an average user that generally accumulates in some number

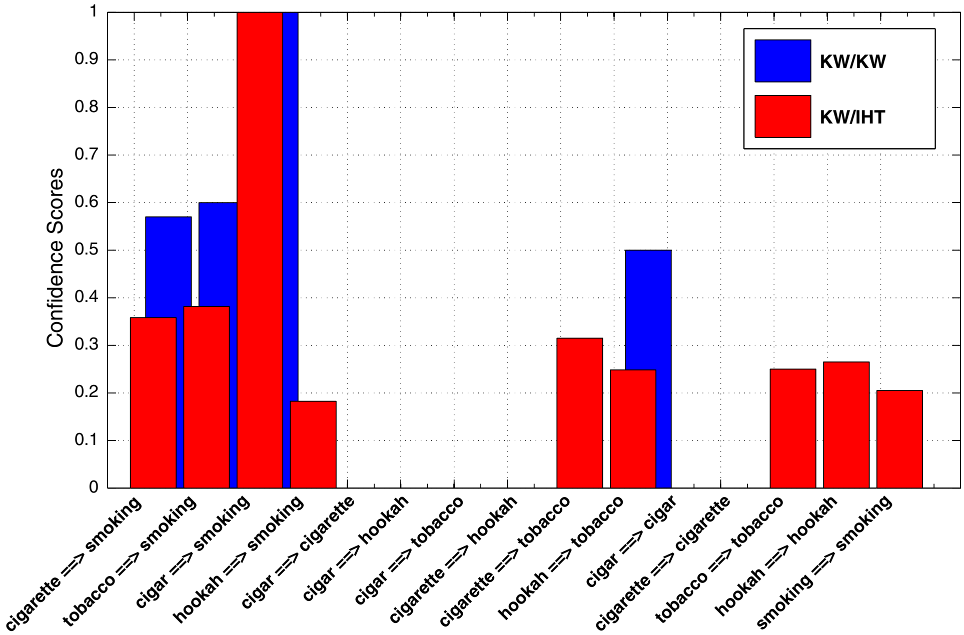


Fig. 7 Analysis of synthetic dataset—rules discovered using identical hashtags experiment

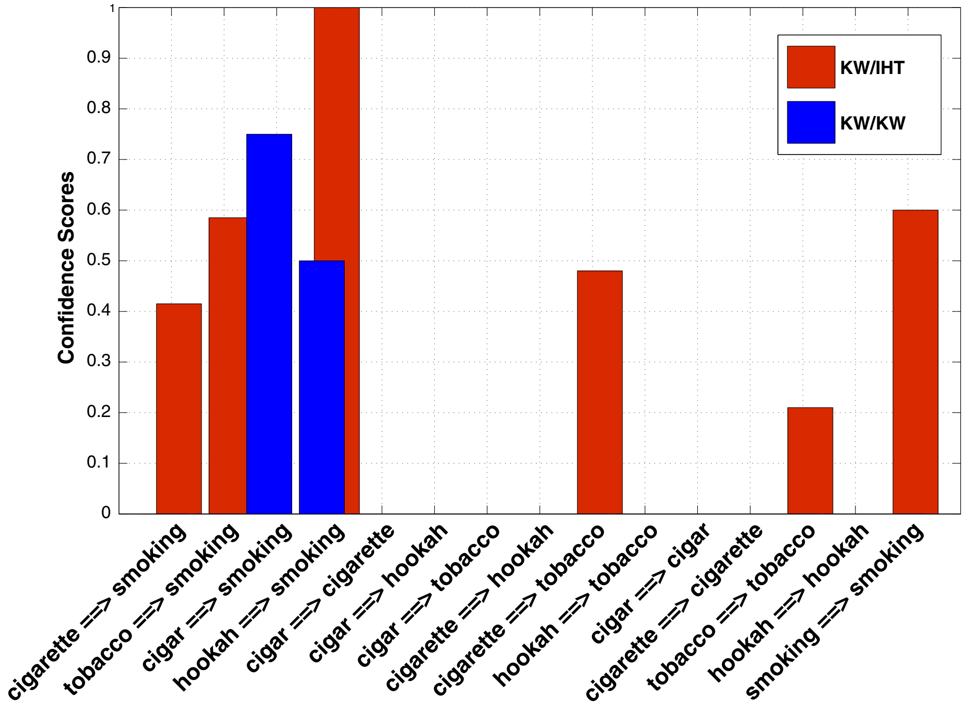


Fig. 8 Analysis of real-tweet dataset—rules discovered using identical hashtags experiment

Table 4 TobaccoQuit account statistics

Item	Tweets	Followers	Friends	Lists
Count	23K	1,695	1,100	9

of years. The system performance is summarized in Table 4.

We also visualized the world-cloud of the entire 23K tweets, which shows consistent results of the pre-selected search keywords that we have chosen for the tobacco quitting campaign along with related topics. Words such as Tobacco, Quitting, Cancer, and Lung are indicated with a bigger size font to indicate their relative importance. However, the visualization also showed surprising results by capturing topics that were not part of the original set of search keywords. Those topics have appeared in smaller font, but we believe there ought to be a correlation between the smoking cessation and these topics (e.g., alcohol, stroke). Additionally, we also grouped followers by a word-cloud topic which shows a great deal of overlapping interests. Figure 9 presents these results. The system received 1,600 different types of impressions from Twitter users. Some of these impressions were: 750 direct replies to tweets we sent, 500 retweets, and the remaining were direct messages to ask questions privately. Most of the impressions received were positive. We also monitored the clicks to the service URL link using Bitly[1] (Incorporation 2013).

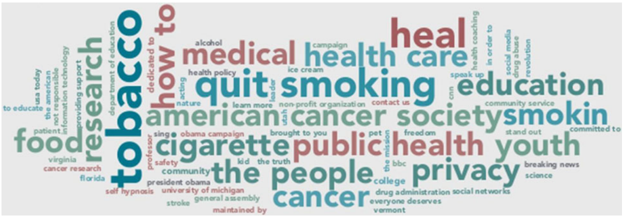


Fig. 9 TobaccoQuit word-cloud visualization

We launched our campaign in May, 2013 and we have received around a 1,000 clicks on the study's website. We found that 81 % of traffic was generated by Twitter by itself. The remaining 19 % of the traffic was generated out of email clients, mobile devices, instant messages and other. Figure 10 shows the daily numbers of clicks for the traffic ratio.

8 Discussions

We have presented an emerging smoking cessation application that is based on Twitter. We are taking an incremental development approach and deploying it in a test environment to obtain the general public's feedback on the services we launch. This approach is intentional, as we must comply to the recruitment protocol proposed in the NIH proposal specification and the IRB board. We proposed two novel algorithms to rank social media users and to develop recruitment services of any type. Our smoking

[1] Real time update: https://bitly.com/ZDMaA7+/global.

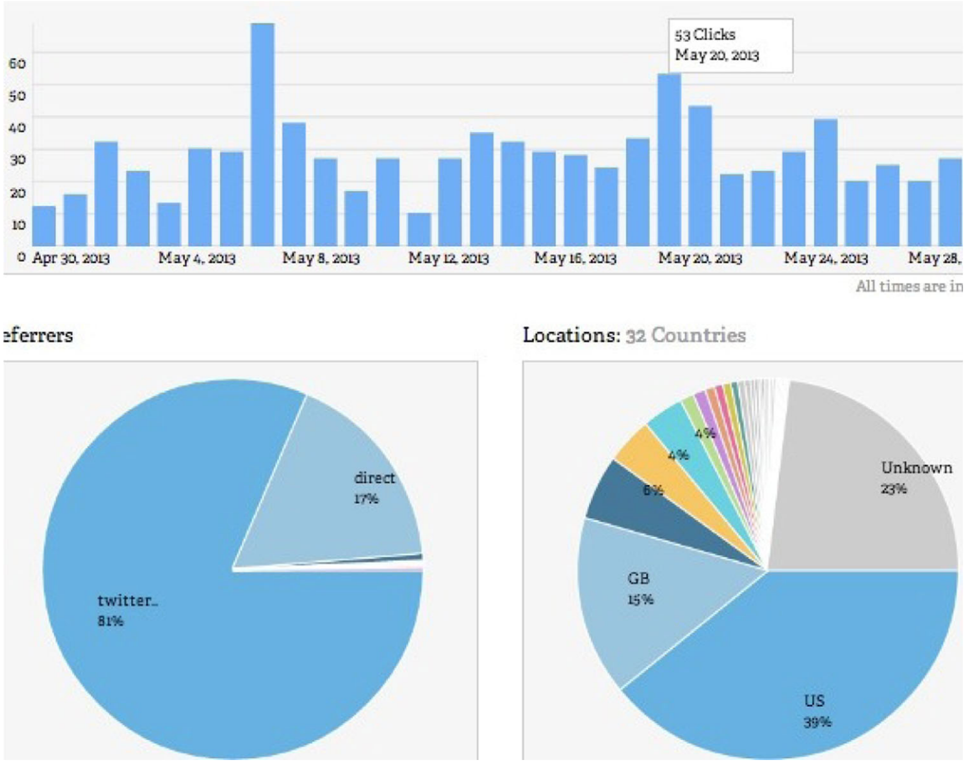


Fig. 10 TobaccoQuit traffic using BitLy

cessation recruitment system can also be extended to launch services for treatment of drug and alcohol problems and other medical recruitment services.

Twitter does not expose the impressions API to the public yet. Therefore, the current software still requires minimal manual labor. We have a designated person that carefully reviews the various impressions and manually responds to general Twitter users or followers. Another limitation is that the system is unable to share external web news beyond the Twitter platform. However, we get around this by using Google Alerts, which delivers new web articles to a designated mailbox. News web links are manually annotated and deposited into the database to be shared.

We have not acquired access to a web link database. This limitation prevents us from computing the in-degree PageRank for links contained in the tweet body. We are resorting to compute pseudo-rank based on the out-degree links. Another limitation to the current algorithms is the lack of adaptability in response to tweet traffic. This would help because the tweet rate varies during the day. Fixed rates could be high during a high traffic time, which would be viewed as spammy. At other times, the services might not be effective if the rates are too low when traffic is low. We hope to address this as we learn our capacity to handle recruitment requests from Twitter users.

From a data mining and knowledge acquisition perspective, the experiments performed on two sets of experiments have not shown a consistent trend especially the ones with lower support (0.01). In both types of experiments we observed fluctuation in the average confidences. Some fluctuation was for the rules that contained identical hashtags others were for the experiments that contained entirely new hashtags. We also believe that the average confidence abstract much of rules specificity which stands in the way of learning which rules are useful for the system. The average measure fails to highlight which rules are significant than others. In order for the system to mature, we must know which rules are significant which could not be achieved using the average confidence measure. Further experiments and measures will be explored in future studies to incorporate hashtags and also computation evaluate the ones produced by the association rule mining experiments such as the ones above. Manual inspections of the association rules that were generated have been found more interesting and worthwhile stating here: (1) the term Marijuana that was highly associated with the following hashtags (#weed, #pot, #grass). Most Twitter users refer to Marijuana but one of those words for a slang. Marijuana was also found associated with the hashtag (#cannabis) which is the plant organism that people use to smoke and inhale the marijuana substances. More interestingly, marijuana was also found highly associated with a hashtags called (#420), after investigating the hashtag, we found out that this was a reference to the 20th April, which is a codeterm used primarily in North America that refers to the consumption of marijuana and by extension, as way to identify oneself with cannabis subculture.

We also presented a case study for a smoking cessation recruitment. Since the development of the early developments of this rule-based system, we have used it for other recruitment studies and beyond nicotine addiction. The most important study is called Geomed Science[2] its purpose is to understand the relationship between the environment around a person's home and their body weight. The investigators suspect that some kinds of environment help people to maintain a lower weight than others.This research may show us ways to support healthier lifestyles. In the same manner of smoking cessation, participants are recruited via means of tweets, favorites, and mentions. There is indeed manual work to prepare the tweets that the system is sending out. Once they are created, they can be diversified. The tweets are designed with place holders to inserting an exact different token types (e.g., geolocation). A tweet such as "Does the #environment in [city] make people fat?" can be instantiated in many different way by replacing the "[city]" placeholder by actual city and state name. Such instantiation is automated and accomplished by means of string manipulation. We acquired geolocations using the gazetteer ontology (Bioportal 2009). As for the rules, they are entirely replaced by rules produced by the experts in the environmental health domain. This demonstrates the strength of developing our systems using an expert system shell environment that can be applied in various domains given sufficient rules communicated by the experts.

Acknowledgments This work is partly funded by National Institute of Health (NIH) Grant: 1 R01 CA165080 and Vermont EPSCoR award number EPS-1101713. We also thank the HPBL team for populating our database of tweets. Special thanks to the Twitter4J project lead, Yusuke Yamamoto, and all contributors. Authors also thank Dr. Benjamin Littenberg for offering another case study to demonstrate the strength and flexibility of the system.

References

Agrawal R, Srikant R (1994) In: Proceedings of the 20th International Conference on very large data bases morgan kaufmann publishers Inc., San Francisco, VLDB '94, pp 487–499. http://dl.acm.org/citation.cfm?id=645920.672836

Aiello A, Burattini E, Tamburrini G (1995) Int J Intell Syst 10(8), 735. doi:10.1002/int.4550100804

B Incorporation (2013) Url shortening and bookmarking services. http://bitly.com/

Breiman L (2001) Mach. Learn. 45(1), 5. doi:10.1023/A:1010933404324

Bioportal N (2009) Gazetteer Ontol. http://bioportal.bioontology.org/ontologies/GAZ

Black A, Mascaro C, Gallagher M, Goggins SP (2012) In: Proceedings of the 17th ACM International Conference on Supporting Group Work, ACM, New York, GROUP '12, pp 229–238. doi:10.1145/2389176.2389211

Chang Y, Dong A, Kolari P, Zhang R, Inagaki Y, Diaz F, Zha H, Liu Y (2013) ACM Trans. Intell. Syst. Technol. 4(1), 4:1. doi:10.1145/2414425.2414429

Chen GM (2011) Comput. Hum. Behav. 27(2), 755. doi:10.1016/j.chb.2010.10.023

Cheong M, Lee VC (2011) Information Systems Frontiers 13(1), 45. doi:10.1007/s10796-010-9273-x

CareerBuilder (1995) http://www.careerbuilder.com

Craigslist (1995) http://www.craigslist.com

Eleta I (2012) In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work Companion, ACM, New York, CSCW '12, pp 363–366. doi:10.1145/2141512.2141621

Fortunato S, Boguñá M, Flammini A, Menczer F (2005) CoRR abs/cs/0511016.

Fortunato S, Boguñá M, Flammini A, Menczer F (2007) Internet Math 4(2):245

Fung FCY (1989) Knowledge-Based Systems 2(4), 228. doi:10.1016/0950-7051(89)90067-1. http://www.sciencedirect.com/science/article/pii/0950705189900671

Ganjisaffar Y (2012) Open source web crawler for java. http://code.google.com/p/crawler4j/

Gayo-Avello D (2013) Soc. Sci. Comput. Rev. 31(6), 649. doi:10.1177/0894439313493979

Ghiassi M, Skinner J, Zimbra D (2013) Expert Syst. Appl. 40(16), 6266. doi:10.1016/j.eswa.2013.05.057

Google (1995) http://www.google.com

Gupta N, Das A, Pandey S, Narayanan VK (2012) In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, KDD '12, pp 1204–1212. doi:10.1145/2339530.2339719

Hamed AA, Wu X (2014) In: Big Data, 2014 IEEE International Conference on (2014)

Hill EF (2003) Jess in action: Java rule-based systems. Manning Publications Co., Greenwich

Hughes JR (2013) J Subst Abuse Treat 45(2), 215. doi:10.1016/j.jsat.2013.01.011. http://www.sciencedirect.com/science/article/pii/S0740547213000342

Hughes JR (2013) Nicotine Tobacco Res 15(2):588. doi:10.1093/ntr/nts154

Huang J, Thornton KM, Efthimiadis EN (2010) In: Proceedings of the 21st ACM Conference on Hypertext and Hypermedia, ACM, New York, HT '10, pp 173–178. doi:10.1145/1810617.1810647

Lee DC, Budney AJ, Brunette MF, Hughes JR, Etter JF, Stanger C (2014) Addict Behav 39(8), 1224. doi:10.1016/j.addbeh.2014.04.010. http://www.sciencedirect.com/science/article/pii/S0306460314001129

Lhotska L, Marik V, Vlcek T (2001) Int J Med Inf 63(1–2), 61. doi:10.1016/S1386-5056(01)00172-1. http://www.sciencedirect.com/science/article/pii/S1386505601001721

LinkedIn (2003) http://www.linkedin.com

Kim M, Park HW (2012) Scientometrics 90(1), 121. doi:10.1007/s11192-011-0508-5

Meng X, Wei F, Liu, X Zhou M, Li S, Wang H(2012) In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York, KDD '12, pp 379–387. doi:10.1145/2339530.2339592

Monster (1999). http://www.monster.com

N.L. of Medicine (2013) Medical subject headings browser. http://www.nlm.nih.gov/mesh/2014/mesh_browser/MBrowser.html

[2] More information on this system can be found at Twitter @GeomedSci and https://redcap.uvm.edu/redcap/surveys/index.php?s=AgCWxtoyMX.

Pavlyshenko B (2013) CoRR abs/1310.3499. http://dblp.uni-trier.de/db/journals/corr/corr1310.html

Pring C (2012) 99 new social media stats for 2012. http://thesocialskinny.com/99-new-social-media-stats-for-2012/

Ravikumar S, Talamadupula K, Balakrishnan R, Kambhampati S (2013) In: AAAI (Late-Breaking Developments), AAAI Workshops, vol. WS-13-17 (AAAI, 2013), AAAI Workshops, vol WS-13-17. http://dblp.uni-trier.de/db/conf/aaai/late2013.html

Stringhini G, Wang G, Egele M, Kruegel C, Vigna G, Zheng H, Zhao BY (2013) In: Proceedings of the 2013 Conference on Internet Measurement Conference, ACM, New York, IMC '13, pp 163–176. doi:10.1145/2504730.2504731

Tsakonas A, Dounias G, Jantzen J, Axer H, Bjerregaard B, von Keyserlingk DG (2004) Artificial Intell Med 32(3), 195. doi:10.1016/j.artmed.2004.02.007. Adaptive Systems and Hybrid Computational Intelligence in Medicine. http://www.sciencedirect.com/science/article/pii/S0933365704001058

Twitter.com (2006) The twitter rest api. https://dev.twitter.com/docs/api

Uzuner O, Goldstein I, Luo Y, Kohane I (2008) J Am Med Inf Assoc 15(1), 14. doi:10.1197/jamia.M2408. http://jamia.bmj.com/content/15/1/14.short

Wang CC, Chien MN, Huang CH, Liu L, (2007) in Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on, vol 4, vol 4, pp 109–115. doi:10.1109/FSKD.2007.117

Yamamoto Y (2007) Java library for the twitter api. http://www.twitter4j.org/

Yan X, Han J (2002) In: Proceedings of the 2002 IEEE International Conference on Data Mining, IEEE Computer Society, Washington DC, ICDM '02, pp 721. http://dl.acm.org/citation.cfm?id=844380.844811

Yang X, Ghoting A, Ruan Y, Parthasarathy S (2012) In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, KDD '12, pp 370–378. doi:10.1145/2339530.2339591

Yang M, Lee JT, Lee SW, Rim HC (2012) In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, New York, SIGIR '12, pp 1073–1074. doi:10.1145/2348283.2348475

Yu K, Ding W, Simovici DA, Wu X(2012) In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, KDD '12, pp 60–68. doi:10.1145/2339530.2339544

Yamaguchi Y, Takahashi T, Amagasa T, Kitagawa H (2010) In: Proceedings of the 11th international conference on Web information systems engineering, Springer, Berlin, WISE'10, pp 240–253. http://dl.acm.org/citation.cfm?id=1991336.1991364

Wu X, Kumar V (2009) The top ten algorithms in data mining, 1st edn. Chapman & Hall, London

Wu X, Kumar V, Ross Quinlan J, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou ZH, Steinbach M, Hand DJ, Steinberg D (2007) Knowl Inf Syst 14(1), 1. doi:10.1007/s10115-007-0114-2

Zhang B, Wang J, Zhang L (2013) In: ICDCS Workshops (IEEE, 2013), pp 190–195. http://dblp.uni-trier.de/db/conf/icdcsw/icdcsw2013.html