Advanced Engineering Technology and Application

ISSN: 2090-9543 VOLUME 4, ISSUES 3, 2015 www.egyptfuture.org/ojs/

# Detection of Diseases using Social Networks and Public Domain Knowledge

*Ramesh Kini*, Aigerim Zinel* and Sabira Arisheva**

Kazakh - British Technical University, Almaty, Kazakhstan

Received: 15 Apr.2015, Revised: 20 Apr. 2015, Accepted: 24 Mar. 2015
Published online: 1 May 2015

**Abstract:** This paper describes how information, taken from social media and public domain knowledge, such as Twitter, can be useful in healthcare and public health management – it describes our proposed technique of: i) collecting tweets with the information about the symptoms users suffer from; ii) filtering them; and iii) applying Dempster-Shafer theory, which deals with uncertainty, for associating the most probable disease with the given symptoms. Additionally, location-related information taken from the tweets or user profiles, using Twitter API, helps health care analysts and planners to identify the regions where the disease could potentially erupt as an epidemic. When this information is superimposed on a geographical map at the local, provincial, national, or global level, to create a heat-map, the resulting GIS tool can help public health specialists, we believe, to arrive at better pre-emptive strategies to tackle such epidemics before they become pandemics, e.g., carry out a selective vaccination program, or a cull of the birds or animals that are the source or vectors of the zoonotic disease, and so on.

**Keywords:** Dempster-Shafer theory, Epidemiology, Social media, Twitter

## 1 Introduction

Mankind has never been as mobile as it is today. Thousands of airplanes and ships travel everyday from one region or nation or continent to another carrying millions of passengers, some of whom may be ill with highly contagious diseases which perhaps have not yet been diagnosed. Consider the fact that scientists at the Center for Infection and Immunity (CII) at Columbia Universitys Mailman School of Public Health have estimated that there is a minimum of 320,000 viruses in mammals awaiting discovery[1], and compare that number with the few hundred viruses at most that healthcare professionals currently do know about and are reasonably well-equipped to tackle. According to lead author of [1] Simon Anthony, ”our whole approach to discovery has historically been altogether too random: What we currently know about viruses is very much biased towards those that have already spilled over into humans or animals and emerged as diseases. But the pool of all viruses in wildlife, including many potential threats to humans, is actually much deeper.

This raises the risks for diseases to spread rapidly around the globe and may well pave the way for the next pandemic that humanity will have to face sooner or later, along the lines of the flu pandemic which killed as many as 50 to 100 million people worldwide in 1918. This requires epidemiologists to discover and track diseases more efficiently, more effectively and more importantly much more rapidly. According to W. Ian Lipkin, director of CII and senior author of the Columbia U. study cited above[1], ”...to provide the viral intelligence needed for the global public health community to anticipate and respond to the continuous challenge of emerging infectious diseases.

According to an epidemiologist at the University of Toronto Dr. David Fisman, ”The way that information moves is very similar to the way disease moves” , and it was words like these that inspired us to try to use social media for tracking diseases in the first place. Our main focus in this research effort has been on using analytical and predictive tools to model and analyze epidemics and to incorporate these tools in our attempt to develop an early warning epidemiological system that could be used to predict and monitor the spread of contagious diseases, based on knowledge available in the public domain.

Social networking services have become an important source of the latest and most-accurate information on current and past events that could have an impact on some

rapidly-evolving situation somewhere around the world. Twitter, as one of the most popular online services, is characterized by short messages related to various domains, and is proving to be a very attractive research topic as well as a rich source of raw data for researchers in a variety of different fields of study. It has been already studied and used for stock-markets [2][3] and election [4] prediction, and as in this paper has also been used to model the spread of influenza across the globe. We have used Dempster-Shafer theory to process and analyze the content shared in Twitter, as the basic raw material, and to produce a heat map of the potential epidemic or pandemic on that basis.

Here, we have sampled and analyzed content from Twitter, retrieved from 10000 users. After eliminating superfluous or irrelevant information, we extract the symptoms and location from each tweet. Dempster-Shafer theory of evidence is then used to map the extracted symptoms to diseases, and to produce a heat-map of the disease as it spreads on a local, national or global basis. This would enable public health analysts or planners to understand how influenza, for instance, is spread across different regions, and to to arrive at better pre-emptive strategies or measures to tackle such epidemics before they become pandemics, e.g., carry out a selective vaccination program, or a cull of the birds or animals that are the source or vectors of zoonotic diseases, such as swine flu, bird flu, ebola, and so on.

The rest of this paper is organized as follows: Section 2 describes how information was gathered from social media. We introduce Dempster-Shafer theory that is used for mapping symptoms to diseases in Section 3, and Section 4 is devoted to the heat-map model and results. The conclusions follow in Section 5.

## 2 Gathering Data from Social Media

Since the data and knowledge we need to collect and analyze is taken from the public domain, we needed a social medium, which users all over the world would post information to very frequently in a short text format, rather that bulky media files, for instance, and which would therefore contain infinitely large amount of data. Thus, Twitter was chosen as the most suitable candidate, given that it offers an enormous pool of data and is the source for the most relevant and updated information, which would be collected and harvested as raw. Another reason is that Twitter has an API, which allows analysts ad users to easily download tweets and analyze them for context and content. Twitter API provides developers easy mining, where data is organized in a nice JSON format . Each tweet has its own ID which is a unique identifier and has a corresponding user, who writes tweets. The information taken from a user profile contains data about the profile settings and also it can be details about the location and language of the user (if enabled) with the following structure:


```
{
"text": "...",
"id": 1,
"coordinates": null,
"geo": null,
"created_at": "Thu Mar 26 05:29:27 +0000
    2015",
"place": null
...
"user":
{
"statuses_count":1,
"favourites_count":1,
"protected":false,
"profile_text_color":"437792",
"profile_image_url":"...",
"name":"Twitter API",
"profile_sidebar_fill_color":"FFFFF",
"listed_count":9252,
"following":true,
"profile_background_tile":false,
"utc_offset":-28800,
"description":"...",
"location":"San Francisco, CA",
"followers_count":665829,
"geo_enabled":true,
"entities":{"hashtags":[],"urls":[]
,"user_mentions":[]}
,"expanded_url":null
,"is_translator":false,
"lang":"en",
"time_zone":"Pacific Time (US & Canada)",
"created_at":"Wed May 23 06:01:13 +0000
},
...
}
```


The data was collected using Twitter API and stored in MySql database. One million tweets collected from 10000 arbitrary users are used as a simple representation of Twitter, which should be enough to analyze and model disease distribution.

The following is pseudocode of our data retrieval algorithm :

```
//List of users to visit is obtained from
    db, or initially is "Obama".
List<User> users = new
    List<User>(){"Obama"};

foreach(User u:users){
try{
if(user.isNotVisited){
List<Status> st = getAllTweets();
insertAllTweetsToDB(st);
}
users.add(user.GetFollowers);

if(getTweetCount=1000000)
```

```
return;
}catch(){
//Wait for n minutes
}
}
```

Twitter API has limits for number of request per window per authorization token. The number of allowed requests depends on the resource you want to access, and the method you are using. We implemented Twitter mining using Java programming language, and had rate limit of 15 minute between each 180 requests. During each request frame our application inserted tweets into MySQL database to store information about user and tweet: such as author, date, location and message.

Some messages retrieved would be ignored during the analysis, such as Non-engish tweets and reposted tweets. Unfortunately, not all tweets have geolocation information, and therefore we had to make some assumptions in order to proceed: we have assumed that the location of the tweet is more likely to be the same as the one from closest tweet and tweets are more likely to correspond the region that is shared in users profile. Thus, only 5 percent of messages were considered to be valid (have location and symptoms) for further analysis using Dempster Shafer theory.

## 3 Dempster-Shafer Theory

Most tasks requiring intelligent behavior have some degree of uncertainty due to unreliable or ambiguous measurements. In health care expert systems the knowledge base is usually defined by the knowledge and experience of doctors. Thereforesome uncertainty management mechanisms are required to handle errors.

The Dempster-Shafer theory was introduced by Dempster [5] and further developed by Shafer[6]. It is mathematical theory of evidence used to calculate probability of an event with belief functions and plausibility reasoning. The Dempster-Shafer theory of evidence has uncertainty management and inference mechanisms analogous to our human reasoning process. The basic representational unit is called a basic probability assignment (bpa) function. The two main operations for manipulating bpa functions are marginalization and Dempsters rule of combination.Building up on previous applications of Dempster-Shafer theory for Skin Diseases Expert Systems [7], Poultry Diseases Warning System [9], African Trypanosomiasis[10] and Insect Diseases[8] Detection and others, we seek to offer a new implementation of Dempster-Shafer theory in health care, and one in particular that uses social media.

Initially, subjective probabilities are assigned to each subset of the frame. Belief in the hypothesis is obtained as the sum of the probabilities of all sets and measures the strength of evidence in a favor of a given hypothesis. It is characterized by a degree of support between 0 (no support) and 1 (certainty) and does not have mathematical properties of probabilities. Plausibility is the sum of the masses of all sets whose intersection with the hypothesis is not empty.

Dempster-Shafer theory provides a function for computing from two pieces of evidence and arriving at the combined influence of these pieces of evidence, by using Dempsters rule of combination to combine their associated masses. In Dempster-Shafer theory a set of all possible answers is called a Frame of discernment, where only one answer is correct. Let m1 and m2 be mass assignments in the frame of discernment. The combined mass is computed using the formula:

$$(m_0 \oplus m_1)(Z) = \sum_{X \cap Y = Z} m_1(Z) m_2(Z)$$

### *3.1 Detecting diseases*

Assuming that we have a finite set of names of all possible symptoms, we search for keywords and parse statuses collected from Twitter. These symptoms will be used in the Dempster-Shafer theory-related calculations later on, so as to correctly map the symptoms to diseases.

The following is the example of the message retrieved from Twitter:

```
Day 2 of #Fever .. #Cough .. #Cold ..
    #Body Ache .. #Sorethroat.. I feel n
    look lika 100 yrz old !!!!! Need to
    get better now .. Lik now now now now !
```

It contains such symptoms as Cough, Bodyache, Fever and Sorethroat. This list of symptoms are used to illustrate an iteration of calculations used to map symptoms to diseases. The table below (Table 1) shows the basic probability assignments (BPA) for each symptom, which is equivalent to a probability mass function in probability theory.

**Table 1:** Basic Probability assignments of each symptom.

| symptom | disease | Condition1, bpa |
|---|---|---|
| Cought | Influenza<br>Cold<br>Bronchitis | 0.62 |
| Bodyache | Influenza<br>Stress<br>Fatigue | 0.41 |
| Fever | Influenza<br>Cold<br>Infection<br>Tonsillitis | 0.57 |
| Sorethroat | Influenza<br>Cold<br>Tonsillitis | 0.51 |

1) Symptom 1 Cough.

Cough is the symptom of Influenza, Cold and Bronchitis with the BPA = 0.62, which also means that the remaining probability is left to the unknown case.

$$m_1\{I,C,B\} = 0.62$$

$$m_1\{\Theta\} = 1 - 0.62 = 0.38$$

2) Symptom 2 - Bodyache.

Bodyache is the symptom of Influenza,Stress,Fatigue with the probability 0.41. However, considering Bodyache and Cough at the same time produces another probabilities whose calculations are shown in the table 2

$$m_2\{I,S,F\} = 0.41$$

$$m_2\{\Theta\} = 1 - 0,41 = 0.59$$

Now we have two symptoms, and therefore should recalculate bpa values for next combinations ($m_3$). Combination rules for the $m_3$ can be seen in the Table 2.

**Table 2:** Combination of Symptom 1 and Symptom 2.

| | | $m_2\{$I,S,F$\}$ | $m_2\{\Theta\}$ |
|---|---|---|---|
| | | 0.41 | 0,59 |
| {I, C, B} | 0.62 | 0.2542 | 0.3658 |
| $\{\Theta\}$ | 0.38 | 0.1558 | 0.2242 |

$$m_2\{I,C,B\} = \frac{0,3658}{1-0,2542} = 0,49$$

$$m_2\{I,S,F\} = \frac{0,1558}{1-0,2542} = 0,208$$

$$m_2\{\Theta\} = \frac{0,2242}{1-0,2542} = 0,3$$

3) Symptom 3 - Fever.

Fever is the symptom of Influenza, Cold, Infection, Tonsillitis with the BPA 0.57. Together with Cough, Bodyache, the probabilities are calculated as follows:

**Table 3:** Combination of Symptom 1, Symptom 2 and Symptom 3.

| | | $m_4\{$B$\}$ | $m_4\{\Theta\}$ |
|---|---|---|---|
| | | 0,57 | 0.43 |
| {I, C, B} | 0.49 | 0.2793 | 0.2107 |
| {I, S, F} | 0.208 | 0.1185 | 0.0894 |
| $\{\Theta\}$ | 0.3 | 0.171 | 0.129 |

$$m_3\{I,C,I\} = \frac{0.171}{1-0.2793-0.1185} = 0.028$$

$$m_3\{I,C,B\} = \frac{0.2107}{1-0.2793-0.1185} = 0.349$$

$$m_3\{I,S,F\} = \frac{0.0894}{1-0.2793-0.1185} = 0.1484$$

$$m_3\{\Theta\} = \frac{0.129}{1-0.2793-0.1185} = 0.214$$

4) Symptom 4 - Sorethroat

Finally, sorethroat is the symptom of Influenza, Cold, Tonsillitis. Considering all four symptoms together gives five probabilities of different outcomes:

**Table 4:** Combination of all symptoms.

| | | $m_4\{$B$\}$ | $m_4\{\Theta\}$ |
|---|---|---|---|
| | | 0,51 | 0.49 |
| {I, C, B} | 0.349 | 0.178 | 0.171 |
| {I, S, F} | 0.1484 | 0.076 | 0.073 |
| {I, C, I} | 0.028 | 0.014 | 0.014 |
| $\{\Theta\}$ | 0.214 | 0.109 | 0.105 |

$$m_4\{I,C,T\} = \frac{0.109}{1-0.178-0.076-0.014} = 0.148$$

$$m_4\{I,C,B\} = \frac{0.171}{1-0.178-0.076-0.014} = 0.233$$

$$m_4\{I,S,F\} = \frac{0.073}{1-0.178-0.076-0.014} = 0.1484$$

$$m_4\{I,C,I\} = \frac{0.014}{1-0.178-0.076-0.014} = 0.1484$$

$$m_4\{\Theta\} = \frac{0.105}{1-0.178-0.076-0.014} = 0.214$$

The final calculation gives us probability assignments for each disease. The disease or list of diseases with the highest bpa are better candidates to be mapped to. Thus, the symptoms 'headache, fever, sore throat, cough' are mapped to the list of diseases {Influenza, Cold, Bronchitis} with the highest bpa 0.233

## 4 HeatMap

We have obtained a set of location and disease pairs that can be used to fill a weighed heatmap as hotspots representing most serious epidemiological cases. We have used Google Maps JavaScript API to create a web page based on our epidemiological model of the situation as it evolves. The following is the heat map generated on the basis of data retrieved from Twitter.

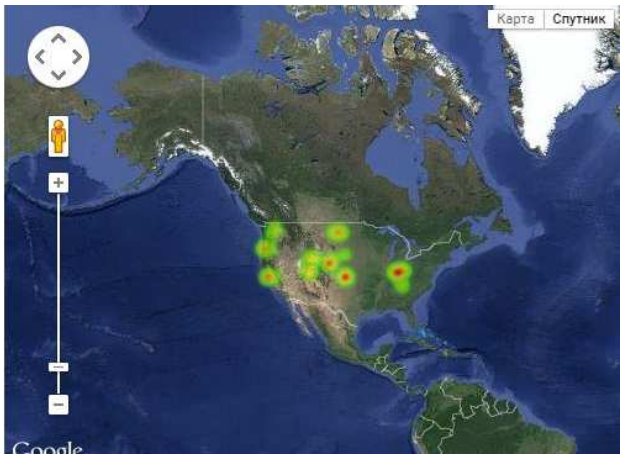


**Fig. 1:** Heatmap.

## 5 Conclusion

Information shared in social media such as Twitter, can be used to bring about a quick and preemptive operational response for any currently evolving epidemiological situation in the world. The application of Dempster-Shafer theory to symptom-related data retrieved from Twitter enables us to map these symptoms to the most probable diseases that people may be suffering from, based on the tweets they post everyday. This paper describes only an analysis of existing tweets. This method is however also useful for analyzing the behaviour of diseases spread in the past, at present and also in the future. That is, if we could extend these predictive capabilities and build effective early warning systems, and if data could be collected and analyzed efficiently as well as quickly, i.e., in near real time, then the model predictions could more accurately reflect the realities on the ground. We envisage that such a system would enable health care authorities to respond more quickly, and perhaps even preemptively, to ongoing events and draw up more effective strategies and plans, e.g., vaccination or quarantine programs or even culling campaigns in the case of birds and animals, etc., to combat zoonotic diseases before they spread too rapidly and over ever-larger geographical regions, and before these diseases can overwhelm the capacities and capabilities of the health care systems that are in place.

## References

[1] Simon J. Anthony, Jonathan H. Epstein, Kris A. Murray, Isamara Navarrete-Macias, Carlos M. Zambrana-Torrelio, Alexander Solovyov, Rafael Ojeda-Flores, Nicole C. Arrigo, Ariful Islam, Shahneaz Ali Khan, Parviez Hosseini, Tiffany L. Bogich, Kevin J. Olival, Maria D. Sanchez-Leon, William B. Karesh, Tracey Goldstein, Stephen P. Luby, Stephen S. Morse,Jonna A. K. Mazet, Peter Daszak, W. Ian Lipkina, ”A strategy to estimate unknown viral diversity in mammals”, *The American Society for Microbiology - mBio open-access online journal*, September 2013

[2] Anshul Mittal, Arpit Goel, ”Stock Prediction Using Twitter Sentiment Analysis”, 2012.

[3] Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, Xiaotie Deng, ”Exploiting Topic Based Twitter Sentiment for Stock Prediction”, *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, 2013

[4] Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe, ”Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment”, in *Fourth International AAAI Conference on Weblogs and Social Media*, 2010

[5] Dempster, A.P., ”A generalization of Bayesian inference”, *Journal of the Royal Statistical Society*, 1968

[6] Shafer, Glenn, ”A Mathematical Theory of Evidence.”, *Princeton University Press*, 1976

[7] Andino Maseleno, Md.Mahmud Hasan, ”Skin Diseases Expert System using Dempster-Shafer Theory, *Intelligent Systems and Application*, May 2012

[8] Andino Maseleno, Md.Mahmud Hasan, ”The Dempster-Shafer Theory Algorithm and its Application to Insect Diseases Detection”, *International Journal of Advanced Science and Technology* Vol.50, January 2013.

[9] Andino Maseleno, Md.Mahmud Hasan, ”Poultry Diseases Warning System using Dempster-Shafer Theory and Web Mapping”, *International Journal of Advanced Research in Artificial Intelligence*, Vol. 1, No. 3, 2012.

[10] Andino Maseleno, Md.Mahmud Hasan, ”African Trypanosomiasis Detection Using Dempsert-Shafer Theory”, *Journal of Emerging Trends in Computing and Information Sciences*, Vol. 3, No. 4, April 2012

**Ramesh Kini** is currently Professor at Kazakh-British Technical University. Received PhD in Carnegie Mellon University - Tepper School of Business in Operations Management. His main interests are analysis, management consulting, developing business strategy, business planning, as well as business analysis, and risk management.

**Aigerim Zinel** received the bachelor degree in Automation and Control at Kazakh-British Technical University in Almaty. Her main research interests are: big data, fuzzy logic, modelling and analysis of contagious diseases.

**Sabira Arisheva** received the bachelor degree in Computer Science and Software at Kazakh-British Technical University in Almaty. Her main interests are: big data, information modelling and web/mobile applications.