

# Informatics Research Proposal

## Predicting influenza trends from blogs

Dan Harvey

### 1 Introduction

Influenza epidemics occur annually all over the world and affect between 5 to 15% of the population [11]. This causes between 250,000 and 500,000 deaths each year [11] which could be reduced with a rapid response to an outbreak with vaccinations. The Centre for Disease Control (CDC) in the US currently compiles a weekly report on influenza activity based upon viral surveillance labs' and doctors' reports [1]. This means that there is a week's lag for the detection of an influenza case to be reported and even longer unknown lag time for the case to appear at a lab or doctors in the first place. With the lag time being so long the influenza epidemic could be well established before any action could be taken to vaccinate people.

This project is looking at the problem of reducing the lag time between cases of influenza occurring and the detection of an outbreak from an indirect data source. Indirectly monitoring sources of data collected and produced for other means brings a lot of benefits as the information is already there, it just needs to be analysed in a novel way. By reducing the lag time between influenza cases occurring and their detection, health centres will be able to react sooner to new epidemics which could potentially save lives.

### 2 Background

Over the past 10 years there has been a growth in research on tracking disease trends using indirect data sources, such as phone calls to help lines [2], sales of drugs [9], and web access logs [5]. Wagner et al. in [10] looked at research done on the detection of disease outbreaks and investigated how the timeliness of detection could be improved. They looked at various signals for detecting influenza which varied from a 5-week lag time to 4 weeks ahead, but there was still a high overhead lag in reporting these. They also suggested

various ways to increase the timeliness, one of which was adding new signals from data collected routinely for other purposes, which I am going to look at in this paper.

Another source of data routinely collected from a wide range of the public is web search logs. These were first looked at in [3] where they indirectly tracked search terms by using clicks on adverts using Google Adwords on Canadian Google search results. The adverts were for searches with ‘flu’ or ‘flu symptoms’ keywords and advertised information about the influenza flu. The click data was fitted to a linear regression model to predict reports from both doctors and labs. They found the data correlated better than reports from doctors but not as well as the lab reports. The lag on their predictions was weekly and could predict a week in advance, this was further than current systems used.

Analysing search logs for trends was researched further in [7] where they used data directly from Yahoo search logs over a 4-year period between 2004 and 2008. They tracked the frequency of searches containing hand-picked terms related to influenza. This was done by using the fraction of searches containing these terms over the total number of searches. The frequency was calculated over a week of search and they estimated the location within the US from IP addresses. A linear regression model was then fitted to the data with a target of the lab reports and influenza-related mortality rate. They found that using a week of searches produced the strongest correlation with data from the labs; this allowed them to predict an increase in lab reports two weeks in advance. They also fitted a model for each of the 9 US census regions and found that even though there was still a correlation it was not statistically relevant enough over all 9 regions to be reliable.

The latest research with search logs was done at Google in [4] where they used 5 years of search logs over the period between 2003 to 2008. Rather than hand picking influenza-related search terms as in [7] they designed an automated method to pick them. This was done by calculating how well each unique search term could predict the number of influenza reports from doctors for each region of the US. The variation of the search terms over the regions was also taken into account to reward terms which showed a stronger variation as it was unlikely all the regions would be changing at the same rate. They found that the top 45 queries produced the best fit so they used these as the parameters of the model. They then fitted a linear model to weekly frequencies of these terms. They tested their model in the 2007 08 flu season and found that each week they could accurately predict the reports from the CDC. Their model could predict the CDC data between 1 to 2 weeks ahead.

This project will aim to improve the lag time between cases of influenza

occurring and the detection of the outbreak. This will be done by indirectly tracking blog posts for influenza related features that can be used to produce a prediction for the trend of influenza. The rapid growth of user-created content on the Internet creates a vast quantity of new data that could contain information from people about their illnesses before they go to the doctors, if they do at all. As social networking sites such as Facebook and Twitter continue to grow, with people posting their lives online, this will become a greater potential to monitor the health of a population indirectly. By utilising more data than can be acquired from search logs this project will try to predict influenza trends with a lag time of less than other sources available today.

### 3 Method

To test my hypothesis for this project I will come up with a regression model to predict the trend of influenza in the US. There are three main parts to doing this; first the raw text from the blogs needs pre-processing into features to describe them; secondly, due to the nature of natural language, the dimension of these features needs reducing to find the ones that are useful in predicting the influenza trend; and finally, a regression line needs to be created from these features which is used to get the predictions.

The raw text from documents will be represented in the bag of words format by splitting the text into word phrases on spaces. This is where each unique word from a document will be put into a set and all order will be lost. This is a good representation as it has been shown in information retrieval [8] and also text classification[6] that it produces the best performance and any more complex models of text do not improve performance. The frequency of words from natural language also follows the Zipf distribution where the frequency of a word is inversely proportional to its rank. This means that there will be many high-frequency words common to many documents, known as stop-words, and also many low-frequency words that will only occur once over millions of documents. These will need filtering out to remove noise for feature selection and regression.

Feature selection and dimensionality reduction is where the main focus of this project will be. The aim of this is to find important features that correlate well to the influenza trends or create new features which are a combination of existing ones. There are many possible ways to do this such as selecting terms which are words related to influenza, or automated methods which compare each individual feature to its ability to predict the regression line. As these depend upon the data set there is no way of knowing which is

best to use, so a wide range of these methods needs to be tried and evaluated. To do this I will create a framework that will allow me to efficiently test and compare many of these feature selection methods. I will evaluate these using methods described in Section 4.

I will use linear regression to produce a trend line from the data, using the CDC weekly data as my target values. A support vector machine (SVM) will be used for the regression as these have been found to perform very well with high feature dimensions and are also quite efficient given their performance.

The data set that I have to work with will be split into three parts which are a training, validation and test set. This will be done by splitting at two dates to create three different time segments. The training set will be used to create an initial regression line, the validation set will then be used to test different feature selection and SVM parameters. Finally the test set will be used only to evaluate final models to report for the project.

## 4 Evaluation and Outcomes

To evaluate the regression model I will compare the trend line produced from the test set to the weekly CDC influenza data. I will do this by calculating the  $R^2$  error between the two sets of values to find how well they correlate. I will also report the lag time of the model for predicting the level of influenza and the time window that is used to calculate the level. These three values will allow me to also compare the model and results to similar work done by researchers at Google and Yahoo.

The main outcome of this project will be a method that can predict the frequency of CDC data for a given time period in advance. This will be in the form of a feature selection method and regression model to produce the trend line. This will show how viable it is to further research ways of processing publicly available information on the Internet to track influenza and potentially other diseases as well.

## References

- [1] Center for Disease Control. Overview of influenza surveillance in the united states, 2009. <http://www.cdc.gov/flu/weekly/pdf/overview.pdf>.
- [2] D.L. Cooper, G.E. Smith, S.J. O'Brien, V.A. Hollyoak, and M. Baker. What can Analysis of Calls to NHS Direct Tell us about the Epidemi-

- ology of Gastrointestinal Infections in the Community? *Journal of Infection*, 46(2):101–105, 2003.
- [3] G. Eysenbach. Infodemiology: Tracking Flu-Related Searches on the Web for Syndromic Surveillance. In *AMIA Annual Symposium Proceedings*, volume 2006, page 244. American Medical Informatics Association, 2006.
- [4] J. Ginsberg, M.H. Mohebbi, R.S. Patel, L. Brammer, M.S. Smolinski, and L. Brilliant. Detecting Influenza Epidemics using Search Engine Query Data. *Nature*, 2008.
- [5] H.A. Johnson, M.M. Wagner, W.R. Hogan, W. Chapman, R.T. Olshewski, J. Dowling, and G. Barnas. Analysis of Web Access Logs for Surveillance of Influenza. *Medinfo*, 11(Pt 2):1202–6, 2004.
- [6] D.D. Lewis. An evaluation of Phrasal and Clustered Representations on a Text Categorization Task. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 37–50. ACM New York, NY, USA, 1992.
- [7] P.M. Polgreen, Y. Chen, D.M. Pennock, and F.D. Nelson. Using Internet Searches for Influenza Surveillance. *Clinical Infectious Diseases*, 47(11), 2008.
- [8] G. Salton and C. Buckley. Term Weighting Approaches in Automatic Text Retrieval. 1987.
- [9] M.M. Wagner, J.M. Robinson, F.C. Tsui, J.U. Espino, and W.R. Hogan. Design of a National Retail Data Monitor for Public Health Surveillance. *Journal of the American Medical Informatics Association*, 10(5):409–418, 2003.
- [10] M.M. Wagner, F.C. Tsui, J.U. Espino, V.M. Dato, D.F. Sittig, R.A. Caruana, L.F. McGinnis, D.W. Deerfield, M.J. Druzdzal, and D.B. Fridsma. The Emerging Science of Very Early Detection of Disease Outbreaks. *JOURNAL OF PUBLIC HEALTH MANAGEMENT AND PRACTICE*, 7(6):51–59, 2001.
- [11] World Health Organisation. Influenza fact sheet, 2003. <http://www.who.int/mediacentre/factsheets/fs211/en/>.