EXTRACTING SIGNALS FROM NEWS STREAMS FOR DISEASE OUTBREAK PREDICTION

Sunandan Chakraborty and Lakshminarayanan Subramanian

Department of Computer Science
New York University
New York, USA
{sunandan,lakshmi}@cs.nyu.edu

ABSTRACT

Emergence of digital news provides new opportunities in information extraction. Proper characterization of unstructured news can help identify signals that may drive variations in many observable phenomena, such as disease outbreaks. In this paper, we propose a method to extract such signals from a large corpus of news events and identify a subset of signals that are closely related to the observed phenomenon. We show how words appearing in a large news corpus can be represented and latent features can be extracted to build predictive models. We build and evaluate such a system specifically for characterizing and predicting diseases outbreaks in India. We focused on 5 different diseases prevalent in India and experiments showed that our model can predict disease outbreaks 2 to 4 weeks prior, with an average precision of around 0.80 and recall of around 0.65. We also compared our model with an LDA-based baseline model, where our model demonstrated around 5–14% improvement across different diseases.

1. INTRODUCTION

Automatic disease surveillance is a major public health challenge. There are many data-driven solutions proposed to address this issue, using different types of data sources, such as web search query logs [1][2][3], media reports [4], social media [5], medical data [6][7], emails [8] and call data [9]. Prior works have two broad categories – visualization tools [4] [12] and forecasting tools [9][10][11]. Most prior work relies on private or proxy data sources (search logs), which either limits their use or makes them hard to generalize; certain solutions are also known to be ineffective beyond a limited time period. For example, using search log data for prediction has shown to overestimate the outbreak rate for a variety of reasons [13]. Finally, most of the methods and datasets used in prior work cater to a single disease and it is difficult to replicate them for other diseases.

In this paper, we present a disease surveillance and prediction system using news streams. The primary advantage of this approach is that the dataset is easily available, making the method more generalizable. The main premise of our method is to identify events from news data that are highly likely to be associated with a particular disease outbreak. Over time such events might change for a disease or these events might be different for different diseases. Our methods are designed to be tolerant to these variations. Our model looks at past news to identify typical events that are associated with a disease outbreak, as well as current news to track emerging events that can potentially explain an outbreak. Hence, this model is dynamic in nature and is unlikely to overestimate outbreaks due to staleness in the data.

Using unstructured news streams for such a purpose has three inherent challenges: high volume, high dimensionality, and noise. This paper addresses these challenges using a combination of matrix factorization techniques combined with labeled examples to derive disease-specific condensed predictive models from news sources. This paper aims to combine two diverse data sources to address the disease outbreak prediction problem – unstructured news data sources and structured time series of different outbreaks at a disease specific granularity while carefully removing the noise and outliers in both datasets. Having very different properties, combining these two types of data efficiently and effectively are some of the main challenges addressed in this paper. We propose a matrix-based representational scheme for both the news and the disease data. The advantage of this representation is that it can handle the large size and the growing nature of the datasets. Also, efficient and distributed algorithms can be performed on large matrices. We propose a supervised factorization of the matrices, combining the news and the disease data. The factors of the matrices produce latent features that can help to extract signals from the news to understand the variation of the disease indicators and use them effectively to build the predictive model.

This paper makes two distinct contributions – (a) capture news streams and process the unstructured text to extract signals and represent these signals in a structured form, (b) connect these signals to a variety of observed phenomena to build predictive models for each phenomenon. We implemented

and evaluated our disease outbreak prediction model using a news corpus containing 7 years of articles. Applying the model to predict outbreaks of 5 diseases prevalent in India, we show that our model can predict with an average precision of around 0.80 and recall of around 0.65.

2. EXTRACTING SIGNAL FROM NEWS

There are many methods proposed to extract signals from news corpora. Most of these methods fall into two broad categories – linear and non-linear (graphical). Linear structures [14][15] – although simple – are not able to represent the complex nature of the data. As a result, many newer systems have moved beyond the conventional list-type output and use graphical structures to represent and analyze news data[16] [17]. Graphical representations are better in dealing with the complexity of the data but fail to generalize when the data is dynamic, growing, and have emerging components.

Unstructured news streams can be assumed to be a continuous flow of information of different types of textual features, such as words (or unigrams), noun phrases, entities, topics, collocated phrases, etc. A large corpus of news data within a time period can be converted into a collection of time-series of these text features. In other words, a large corpus of news articles can be represented as a matrix, where each row represents a time-series. Suppose, there are *V* unique words in a corpus of news articles, i.e. the vocabulary and *T* represents the time (in weeks, days or even hours) of the time span of the corpus, i.e. the news articles in the corpus ranges from time 1 to *T*. Then the entire corpus of articles can be represented as a matrix $\mathbf{X} \in \mathbb{R}^{V \times T}$, where the element at i^{th} row and j^{th} column, Θ_{ij} represents a weight (e.g. frequency) of the word i at time j. If this weight is normalized then Θ_{ij} will be ranged between $\{0, 1\}$. Say, $\mathbf{W} \in \mathbb{R}^{K \times V}$ and $\mathbf{Z} \in \mathbb{R}^{K \times T}$ are two matrices such that $X \approx W^T Z$, then $\mathbf{W}$ represents the distribution of the words over a latent feature and $\mathbf{Z}$ represents the latent features distribution over time. We use this K-dimensional latent feature as the signal extracted from news over time.

This is a general representation to identify of latent features for any large news corpus. The main objective of this work is to understand the factors that are influencing or being influenced by a phenomenon. Let $y_{t=1}^{T}$ be the phenomenon we wish to track over the time period $t_1, t_2, ..., t_T$. y can be any observable phenomenon but in this paper we solely concentrate on outbreaks of specific diseases. Our goal is to identify $\mathbf{Z}$, the latent features that are more likely to be associated with phenomenon y and track the likelihood of these features at time t to build the predictive model for y. So, the decomposition of the original matrix $\mathbf{X}$ is done based on the variation of y within the time period $1, 2, ..., T$. So, the matrix decomposition we propose in this paper is a supervised matrix factorization approach [18], where the factors, $\mathbf{W}$ and $\mathbf{Z}$ are dependent upon the phenomenon of interest y. Thus for different y, this approach would produce different factors. So, the matrix factorization of the news corpus matrix $\mathbf{X}$ is defined as a function $g^y(\mathbf{X}) \rightarrow \mathbf{W} \times \mathbf{Z}$, with the parameter $K^{(y)}$ representing the dimension of the latent features and $K^{(y)} \ll N$.

In this paper, we assume that y is a binary variable, where $y_t = 1$ represents the phenomenon is observed at time t. In the case of disease surveillance, $y_t = 1$ means there was an outbreak of disease y at time t. The non-negative matrix factorization of the matrix $\mathbf{X}$ is centered around the variable y. The aim of the factorization is to find the latent features in such a way that we find separate sets of features (ω^+) that are more likely to be associated with the positive examples and similarly features likely to be associated with negative examples (ω^-).

2.1. Factorizing the Matrix with Labeled Examples

The factorization problem in hand is to find two factors of $\mathbf{X}$ (entire news data matrix), $\mathbf{W}$ and $\mathbf{Z}$, with dimensions $N \times K$ and $K \times T$ respectively. The added constraint to the factorization is that there is an additional vector $\mathbf{Y}$ denoting the external phenomenon, in our case outbreak of a certain disease. $\mathbf{Y}$ is represented as a $2 \times T$ matrix, representing the state of the disease for the entire time period $(t_1, t_2, ..., t_T)$ and each $\nu_i \in \mathbf{Y}$ is a binary one-dimensional vector where $\nu_{i0} = 1$ if value of $y_{ti} = 0$ and $\nu_{i1} = 1$ if value of $y_{ti} = 1$. We consider a joint factorization of the original matrix $\mathbf{X}$ and the disease vector $\mathbf{Y}$, where $\mathbf{Y}$ is combined with the second factor $\mathbf{Z}$, so that the disease vector gets associated with the time distribution of the latent factors. Hence, the loss function for this factorization becomes,

$$\begin{aligned} \mathcal{L} = \underset{\mathbf{W,Z,U}}{\operatorname{argmin}} \Big[& ||\mathbf{X} - \mathbf{W}^T\mathbf{Z}||^2 \\ & + ||\mathbf{Y} - \mathbf{U}^T\mathbf{Z}||^2 \\ & + \alpha||\mathbf{W}|| + \beta||\mathbf{Z}|| + \gamma||\mathbf{U}|| \Big] \end{aligned} \tag{1}$$

where, the last three terms are regularization terms with the parameters α, β, and γ respectively and all the factors $\mathbf{W}, \mathbf{Z}, \mathbf{U}$ are non-negative. $l_1$norm based regularization is chosen for sparsity in the latent features [19]. Optimizing $\mathcal{L}$ will produce 3 matrices with the latent features. $\mathbf{U}$ is going to be a $2 \times K$ dimensional matrix representing the latent features' distribution over the observed phenomenon $\mathbf{Y}$, i.e. the distribution of the K latent features when $y_i = 1$ as well as $y_i = 0$ for all $y_i \in Y$. This information is used to build the predictive model for disease outbreak using news events.

2.2. Prediction

Following the factorization of the matrix $\mathbf{X}$, the first factor ($\mathbf{W}$) represent the distribution of the words over the latent features. The second factor $\mathbf{Z}$ is time-series representation of the

latent features. **Z** has a lower dimension than the original matrix, hence, the number of features are also lower. We use this set of latent features to fit a model to predict the value of y at time t. In this paper, we assume $y(t)$ is a binary variable, so we use logistic regression to predict the value of $y(t)$ at time t. In reality, news events might have a delayed effect on y. There might be a time delay δ between the occurring event and its effect on y. To interpret $y(t)$ as being tracked by historical occurrences of news events, we use a parameter δ and look at features occurring at times $t-(\delta+1), ..., t-1$. We fit a model on $x_{i,t-j}$ that best approximates $y(t)$. With the K features at any week t we fit a model to predict $y(t)$ based on logistic regression. The probability of $y(t)=1$ is given by,

$$\mathcal{F}(z) = \frac{1}{1+e^{-\theta^T z}} \tag{2}$$

where, $\theta^T z$ is given by,

$$\theta_0 + \sum_{i=0}^{K}\sum_{j=0}^{\delta} \theta_{i,j} z_{i,t-j} + \epsilon_t \tag{3}$$

where, K is the number of latent features, **z** are the latent features and ϵ is the the error term with a logistic distribution.

3. DISEASE SURVEILLANCE

3.1. Related Work

Preventing large-scale outbreaks of diseases like dengue, malaria, and flu constitute an enormous public health challenge, especially in countries with limited infrastructure committed for prevention, spreading awareness, and containment of these diseases. There are many solutions proposed to mitigate this problem. However, most of these solutions involve manual data collection and analysis. An alternative approach is to use an automated or semi-automated solution, leveraging the ever-expanding ubiquity of the Internet (especially in developed countries). Google Flu trends [12] is one such system that could associate spikes in web searches related to a disease with the actual outbreak of the disease. While that seemed to work initially, over the years such an approach failed to produce the same results as increased awareness of diseases led users to search without the occurrence of the disease, leading to overestimation of flu cases [13, 20]. Moreover, such an approach will not work in regions with very low densities of Internet users. Aggregated search queries from such regions might lead to inaccurate predictions. HealthMap [23] is another disease surveillance system, where news articles about diseases are displayed on a geographical map. However, HealthMap does not provide any analysis on the data and is an information dissemination system. Other similar systems include ProMED-Mail [21] and GPHIN [11]. These have some improvements over the manually operated system described above, but still rely on human intervention in the form of experts supervising data collection, analysis, and reporting. Some recent work has focused on taking this further to build completely automated web-based disease surveillance systems [4, 8, 22]. These systems parse news articles from around the world using sources such as Google News and RSS feeds, as well as social media such as Twitter, to filter and classify articles based on the nature of epidemic, location and news sources. The main drawback of these systems is that they are aimed at identifying information related to diseases and their outbreaks from different sources and put them together as information dissemination or surveillance tool. None of these systems have any predictive modules that can actually predict an outbreak in the near future. Rehman et al [9] have presented a work that can actually forecast dengue outbreaks few days in advance using call data from a local hotline service. While this work has presented accurate prediction results, their approach is difficult to generalize as similar data is not available in other regions and for other diseases. In this paper, we present a model to predict outbreaks for numerous diseases in India using online news stream, which tries to address the limitations in the existing systems.

3.2. Data

We used a corpus of seven years of news articles that has been extracted from the archives of a leading English newspaper from India. There were two goals for this work – (a) identify signals from news streams that are likely to be associated with disease outbreaks, and (b) predict outbreaks of diseases using the signals learned from news streams. We targeted five diseases prevalent in India – flu, malaria, dengue, diarrhea, and tuberculosis (TB). Our training and testing data for the disease outbreak model includes data from 2006 till 2012, from a variety of sources, such as World Health Organization (WHO), Centers for Disease Control and Prevention (CDC), Public Health Foundation of India, and Ministry of Health, Government of India.

3.3. Results

We applied our model presented in Section 2, using the positive examples of outbreaks of the diseases mentioned above. We implemented five different disease outbreak models for each of the diseases. Based on the supervised matrix factorization, we identified the latent features for each of them. As discussed in Section 2, for each disease, we had different factors of the same news matrix **X**. The first factor $\mathbf{W}_{disease}$ represents the word distribution for each of these latent features for a particular disease. We present the top 10 words extracted from the factor **W** for each disease, based on the words' probability. These words are presented in Table 2.

The words found for each disease, shown in Table 2, are a list of automatically identified words that have a strong relationship to the disease. This is an important result as it expands on the key terms related to a disease. This list of

Table 1. Performance for Disease Outbreak Prediction

	Our Model				LDA			
	K=20		K=50		K=20		K=50	
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Dengue	0.835	0.625	0.841	0.638	0.710	0.539	0.747	0.569
Flu	0.793	0.585	0.797	0.603	0.708	0.533	0.722	0.625
Malaria	0.812	0.685	0.801	0.635	0.753	0.568	0.761	0.608
Diarrhea	0.772	0.591	0.796	0.554	0.722	0.545	0.749	0.592
TB	0.793	0.695	0.805	0.688	0.751	0.552	0.774	0.583

Table 2. Disease and related words

Disease	Related words
Flu	H1N1, virus, influenza, swine, pandemic infection, pandemic, avian, symptom, chikungunya
Malaria	leptospirosis, infection, monsoon, falciparum, flood fever, epidemic, encephalitis, bacterial, mosquito
Dengue	leptospirosis, pandemic, mosquito, water chikungunya, chandipura, typhoid, infections, net
Diarrhea	dysentery, biliary, colitis, gastritis,prawns enteric, aneamia, nausea, river, eczema, kebabs
TB	tuberculosis,infection, pandemic, multidrug, flu resistant, virus, vaccine, venereal, hepatitis

words can be used independently to understand how a disease emerges and progresses. For example, only searching for *flu* might not fetch enough results about flu. However, using the expanded list of words (Row 1 in Table 2), such as, *H1N1, swine*, etc., one can have a better resulting articles from a search. For every targeted disease, we used the these words and used Eq 3 to build disease specific predictive models. We trained the model using the data from 2006 till 2010 and tested on the last 2 years of data (2011 and 2012). The results are presented in Table 1.

We implemented two different models by varying the number of latent features (K). We experimented with $K = 20$ and $K = 50$. For each experiment we report the precision-recall values for all the 5 disease outbreak models. Malaria and dengue prediction models have demonstrated better performance. A plausible explanation for this observation is that they are more common diseases. The performance of the models are dependent upon the coverage of its outbreak in the media. These two diseases have more frequent outbreaks and have more drastic impact (such as, mortality). Hence, these diseases are covered more in the news media.

However, there is not much difference in the accuracy when K is varied. $K = 50$ has displayed slight improvement but without further experiments, the exact trend is hard to interpret. We also observe that, generally the recall values are lower than the precision values. This stark difference might be due to only using positive examples in the supervised decomposition of X. Further experiments including the negative examples might boost the recall values.

Comparison with Topic Model: We developed an LDA-based baseline model to compare our method. LDA style topic model is an alternative way of extracting information from text. We used the topics extracted from weekly news articles and their posterior probabilities during each week as the textual feature At time t, the topic distribution of the K topics are computed as – for each $k \in K$ and each document appearing in time t.

$$\phi_t^k := \max_{d \in D_t} \phi_d^k \tag{4}$$

where, ϕ_d^k is the topic proportion of k^{th} topic in document d and D_t represents all documents appearing at time t. We replace the term z with ϕ in Eq 3 to build the LDA based predictive model for disease outbreak. The results are shown in Table 1. The results show that the accuracy is consistently lower for the LDA model. This is due to the inherent nature of news streams and the way LDA computes topics. Although LDA works well for closed domain corpora such as scientific journals, the topic computed for an open-ended corpus like news articles are not well defined. As a result the textual features for disease outbreaks were not robust enough to predict the outbreaks accurately.

More experiments can be performed to identify optimal values for different parameters (e.g. δ in Eq 3). In the present set of experiments, δ was chosen as 4 weeks. Also, in the present setup we assume that the features used in the prediction (Table 2) are independent to each other. A future direction of this work is to construct a graph of disease specific features, using which a more realistic independence assumption can be made.

4. CONCLUSION

In this paper, we proposed a framework to extract signals from unstructured news data and build predictive models for an observed phenomenon using these signals, using supervised non-negative matrix factorization to convert news data into a lower dimensional representation. This paper has solely focused on disease outbreak prediction, the idea is more general and can be extended to build predictive models for other phenomena. Among many possible future directions to this work, one possibility is to extend the use of bigrams, noun phrases, topics, etc. as features. In addition, adding negative samples and learning negative features can potentially improve the prediction accuracy.

References

[1] Emily H. Chan, Vikram Sahai, Corrie Conrad, and John S. Brownstein, Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance, PLoS Negl Trop Dis 5 (2011), no. 5, e1206.

[2] Eysenbach G and Khler C, Health-related searches on the internet, JAMA 291 (2004), no. 24, 2946.

[3] Ginsberg J, Mohebbi MH, Patel RS, Brammer L, Smolinski MS, and Brilliant L, Detecting influenza epidemics using search engine query data, Nature 457 (2009), no. 5, 10121014.

[4] Freifeld CC, Mandl KD, Reis BY, and Brownstein JS, Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports, J Am Med Inform Assoc 15 (2008), no. 2, 150157.

[5] Janaina Gomide, Adriano Veloso, Wagner Meira, Virgilio Almeida, Fabricio Benevenuto, Fernanda Ferraz, and Mauro Teixeira, Dengue surveillance based on a computational model of spatio-temporal locality of twitter., Proceedings of the ACM WebSci11, June 14-17 2011, Koblenz, Germany., 2011.

[6] Das D, Metzger K, Heffernan R, Balter S, Weiss D, and et al., Monitoring over-the-counter medication sales for early detection of disease outbreaks U new york city, MMWR Morb Mortal Wkly Rep 54 (2005), no. suppl, 41U 46.

[7] Bork KH, Klein BM, Mlbak K, Trautner S, Pedersen UB, and Heegaard E, Surveillance of ambulance dispatch data as a tool for early warning, Euro Surveill. 11 (2006), no. 12, 229233.

[8] Tolentino H, "scanning the emerging infectious diseases horizon-visualizing promed emails using epispider", International Society for Disease Surveillance Annual Conference, 2006.

[9] Nabeel Abdur Rehman, Shankar Kalyanaraman, Talal Ahmad, Fahad Pervaiz, Umar Saif, and Lakshminarayanan Subramanian, "Fine-grained dengue forecasting using telephone triage services," *Science Advances*, vol. 2, no. 7, pp. e1501215, 2016.

[10] Eysenbach G, Infodemiology and infoveillance: framework for an emerging set of public health informatics methods to analyze search, communication and publication behavior on the internet, J Med Internet Res 11 (2009), no. 1, e11.

[11] Mykhalovskiy E and Weir L, The global public health intelligence network and early warning outbreak detection: a canadian contribution to global public health, Canadian Journal of Public Health 97 (2006), no. 1, 4244.

[12] Jeremy Ginsberg, Matthew H Mohebbi, Rajan S Patel, Lynnette Brammer, Mark S Smolinski, and Larry Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, no. 7232, pp. 1012–1014, 2009.

[13] David Lazer, Ryan Kennedy, Gary King, and Alessandro Vespignani, "The parable of google flu: traps in big data analysis," *Science*, vol. 343, no. 6176, pp. 1203–1205, 2014.

[14] James Allan, Rahul Gupta, and Vikas Khandelwal, "Temporal summaries of new topics," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA, 2001, SIGIR '01, pp. 10–18, ACM.

[15] Russell Swan and David Jensen, "Timemines: Constructing timelines with statistical models of word usage," .

[16] Qiaozhu Mei and ChengXiang Zhai, "Discovering evolutionary theme patterns from text: An exploration of temporal text mining," in *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, New York, NY, USA, 2005, KDD '05, pp. 198–207, ACM.

[17] Dafna Shahaf and Carlos Guestrin, "Connecting the dots between news articles," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, 2010, KDD '10, pp. 623–632, ACM.

[18] Youngmin Cho and Lawrence K Saul, "Nonnegative matrix factorization for semi-supervised dimensionality reduction," *arXiv preprint arXiv:1112.3714*, 2011.

[19] David M. Blei, Andrew Y. Ng, and Michael I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[20] Donald R Olson, Kevin J Konty, Marc Paladini, Cecile Viboud, and Lone Simonsen, "Reassessing google flu trends data for detection of seasonal and pandemic influenza: a comparative epidemiological study at three geographic scales," *PLoS Comput Biol*, vol. 9, no. 10, pp. e1003256, 2013.

[21] Lawrence C. Madoff, Promed-mail: An early warning system for emerging diseases, Clinical Infectious Diseases 39 (2004), no. 2, 227232.

[22] Collier N, Doan S, Kawazoe A, Goodwin RM, Conway M, and et al., Biocaster: Detecting public health rumors with a web-based text mining system, Bioinformatics 24 (2008).

[23] Clark C Freifeld, Kenneth D Mandl, Ben Y Reis, and John S Brownstein, "Healthmap: global infectious disease monitoring through automated classification and visualization of internet media reports," *Journal of the American Medical Informatics Association*, vol. 15, no. 2, pp. 150–157, 2008.