Unsupervised Public Health Event Detection for Epidemic Intelligence

Marco Fisichella
Forschungszentrum L3S
Hannover 30167
Germany
fisichella@L3S.de

Avaré Stewart
Forschungszentrum L3S
Hannover 30167
Germany
stewart@L3S.de

Kerstin Denecke
Forschungszentrum L3S
Hannover 30167
Germany
denecke@L3S.de

Wolfgang Nejdl
Forschungszentrum L3S
Hannover 30167
Germany
nejdl@L3S.de

ABSTRACT

Recent pandemics such as Swine Flu have caused concern for public health officials. Given the ever increasing pace at which infectious diseases can spread globally, officials must be prepared to react sooner and with greater epidemic intelligence gathering capabilities. However, state-of-the-art systems for Epidemic Intelligence have not kept the pace with the growing need for more robust public health event detection. In this paper, we propose a game-changing approach where public health events are detected in an unsupervised manner. We address the problems associated with adapting an unsupervised learner to the medical domain and in doing so, propose an approach which combines aspects from different feature-based event detection methods. We evaluate our approach with a real world dataset with respect to the quality of article clusters. Our results show that we are able to achieve a precision of 66% and a recall of 81% when evaluated using manually annotated, real-world data. This shows promising results for the use of such techniques in this new problem setting.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Retrospective medical event detection, Clustering, Epidemic Intelligence

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

1. INTRODUCTION

Many factors in today's changing society such as demographic change, globalization, terrorism, as well as the resilient nature of viruses, contribute towards the continuous emergence of infectious diseases. Events such as emerging infectious diseases, are those considered to be either completely new or reoccurring. An important strategy used by officials to mitigate the impact of potential threats, is to find ways to detect the signs of a public health event as early as possible. The body of work devoted to this effort is known as Event-Based Epidemic Intelligence (EI) [11].

In order to provide information as timely as possible, by now all stages of the Event-Based system, including document collection, filtering and processing, are done with little, or no human intervention, using the unstructured and informal text of Web documents to detect facts about current infectious disease activity within a population [3] .

In our proposed solution, we apply the methods of unsupervised event detection in this new problem setting. In doing so, we adapt the unsupervised approach to our problem domain.

We combine burst function analysis with the entity-centric feature representation in a generative model for predictive event detection. Going beyond a random initialization of the probabilities in this generative process, we instead exploit a known distribution of the features that are obtained directly from the burst function.

1.1 Contributions and Outline

The contributions of this work are the:

- Use of an approach to unsupervised event detection and its adaption for the domain of Epidemic Intelligence.
- Incorporation of feature analysis into predictive event detection, by exploiting burst distributions.

The remainder of this paper is organized as follows: in Section 2, we present the details of our approach to unsupervised public health event detection, characterizing the nature of event detection in the public health domain, to lay the foundation for describing the task-specific adaptations required in this setting. Experimental results for our

approach are given in Section 3. Related work is presented in Section 4. Finally in Section 5, we conclude our work.

2. UNSUPERVISED PUBLIC HEALTH EVENT DETECTION

A public health event (PHE) is defined as a specific infection, disease or death that happens at a specific time and place, which may be consecutively reported by many medical articles in a period. The goal of this work is to detect PHEs in an unsupervised manner.

In a typical Epidemic Investigation task, public health officials must detect anomalous behavior. They periodically compute statistics about disease reporting events, using the recent past, in order to build a predictive model for the near future. The model is used as a baseline for detecting any anomalies. These statistics are based on aggregated information, which in our case, is derived from detecting events in an unsupervised manner, from documents on the Web.

We consider a three stage process for detecting unsupervised public health events. In the first stage, Named Entity Feature Representation, we build entity-centric document surrogates that are suitable for the medical domain. The manner in which we extract features and represent documents is outlined in Section 2.1.

The resulting set of features is then used as input for the Unsupervised Public Health Event Detection stage. Each of these stages are discussed in the sections that follow.

2.1 Named Entity Feature Representation

As a first step, we process raw text to build an entity-centric feature representation of the document. Given a collection of text documents, we define a finite set of articles, $\mathcal{A}$ as well as a Health Event Template, $\mathcal{T}$. The template $\mathcal{T}$ represents a set of feature types, which are important for describing public health events. More specifically, we describe a public health event by four attributes that provide information on *who* (victims) was infected by *what* (diseases), *where* (locations) and *when* (time, defined as the period between the first relevant article and the last relevant one). Thus, the template is instantiated as: $\mathcal{T} = < Victim, Disease, Location, Time >$. Instances of the template elements are represented as $< v, d, l, t >$, for Victim, Disease, Location and Time, respectively.

Furthermore, a medical article also can be represented by victims, diseases, locations and time (represented by a discrete value, i.e. the timestamp).

In order to simplify our model, we assume the four kinds of information of a medical article are independent. Thus, the probability of an article is given by the product of the following individual probabilities:

$$p(article) = p(victims)p(diseases)p(locations)p(time)$$

2.2 Feature Analysis

In this approach, features are classified with respect to their periodicity (P_w) and their dominant power spectrum (S_w). The periodicity of a feature refers to its frequency of appearances. If the feature is *aperiodic*, then it occurs once within the period P, and its P_w has a value equal to the period itself. If the feature is *periodic*, then it happens regularly with a fixed periodicity, i.e., $P_w \leq \lceil P/2 \rceil$. The periodicity is a function of the dominant power spectrum which is computed via the discrete Fourier transform applied to the feature distributions, for more details refer to [7].

The dominant power spectrum, S_w, of a feature w is a strong indicator of its activeness at the specified frequency; the higher the S_w, the more likely it is for the feature to be relevant within the dataset.

2.2.1 Identifying Burst for Aperiodic Features

Let $y_w(t)$ be the distribution of the feature w over the time t of the period under observation; further, let $y_w(t)$ be computed as in [7]. Then, for each *aperiodic* feature, we keep only the bursty period, which is modeled by a Gaussian distribution.

$$f_{ap}(y_w(t)) = \frac{1}{\sqrt{2\pi\sigma_w^2}} * e^{-\frac{1}{2\sigma_w^2}(y_w(t)-\mu_w)^2} \quad (1)$$

The well known Expectation Maximization (EM) algorithm is used to compute the Gaussian density parameters μ_k and σ_k [2].

2.2.2 Identifying Bursts for Periodic Features

To model the periodic features we chose a mixture of K Gaussian distributions, where $K = \lfloor P/P_w \rfloor$. The mixture is described as follows:

$$f_p(y_w(t)) = \sum_{k=1}^{K} \alpha_k * \frac{1}{\sqrt{2\pi\sigma_w^2}} * e^{-\frac{1}{2\sigma_w^2}(y_w(t)-\mu_w)^2} \quad (2)$$

for the mixture proportions α_k of assigning y_w into the k^{th} Gaussian distribution.

$$0 \leq \alpha_k \leq 1 \text{where} \sum_{k=1}^{K} \alpha_k = 1, \forall k \in [1, K] \quad (3)$$

The Expectation Maximization (EM) algorithm is used to compute the mixing proportions α_k, as well as the individual Gaussian density parameters μ_k and σ_k [2].

2.3 Detecting Health Events

A core step, in the unsupervised detection of events is the clustering of the articles and generation of events. Formally, from this stage we determine that a public health event has occurred, or is currently occurring.

2.3.1 Approach

Numerous techniques exist for detecting events in an unsupervised way (see Section 4.2). In this work, we choose to do so using retrospective event detection, since it is important in EI to use data historical collection, in order to build a predictive model of public health events for the near future. Additionally, we have chosen a probabilistic generative model for event detection, because they have been proven to be a more unified framework for handling the multiple modalities (i.e. time, content, entity types) of an article and its content.

For these reasons, we base our unsupervised event detection algorithm on the Retrospective Event Detection (RED) [10] algorithm. They rely on a generative model where the articles are produced using multinomial distributions over the feature types. These articles are used later as starting points

for the EM algorithm. In addition, in their work, the multinomial distributions are initialized with random probabilities, thus the generated articles are randomly picked.

As part of our approach, we refine the RED algorithm by going beyond this random initialization of probabilities - exploiting the feature distributions from our Feature Analysis stage. The underlying intuituion for our approach is based on proven results, which show that an initial starting point estimated in a better-than-random way can, in fact, be expected to converge closer to the optimum, than an intial point that is picked at random [14]. In our approach, we aggregate the computed feature distributions over the articles, and use this information into the multinomial distributions of the generative model. Thus the generated articles, used as starting points by EM algorithm, are not totally randomly picked.

Although [8, 12] have proven that the events retrieved are not influenced by the starting points, their EM algorithm needs to be restarted several times, with several different random starting points in order to get a good approximation of events. Supported by the analysis in [14], we do not need multiple restarts of the EM algorithm, since an initial starting point estimated in this way, can be expected to be closer to the optimum than a randomly picked initial point.

2.3.2 Generative Model for Health Events

Our generative model is described in Algorithm 1.

Algorithm 1: The generative model for unsupervised Event Detection

begin
- Choose an event $e_j \sim$ Multinomial(θ_j);
- Generate a medical article $a_i \sim p(a_i|e_j)$;
- Draw a timestamp $time_i \sim N(\mu_j, \sigma_j)$;
- **for** *each feature of it, according to the type of current feature* **do**
  - Choose a $victim_{iv} \sim$ Multinomial($\theta_p|time_i$);
  - Choose a $disease_{id} \sim$ Multinomial($\theta_d|time_i$);
  - Choose a $location_{il} \sim$ Multinomial($\theta_l|time_i$);

end

In the algorithm, the vector θ_j are *event* probabilities initially instantiated at random (here the definition of *event* is according to the formalization of the multinomial distribution); $p(a_i|e_j)$ is the probability for an article, a_i, given an event, e_j; μ_j and σ_j are parameters of the conditional Gaussian distribution given event e_j; θ_p, θ_d, θ_d are vectors of probabilities computed aggregating the feature burst distributions over the $time_i$ of the given event e_j.

3. EXPERIMENTS

In order to analyze the results of the introduced method, we ran several experiments. For the specific task considered here which is public health event detection, no annotated data set is available. Anyway, we performed some analysis on a real-world data set. In this section, the data set used for the experiments is introduced together with the experimental settings and results.

3.1 Dataset

To build our collection, we collected the source documents from the *url* column of the PULS online fact base [6], a state-of-the-art *Event-Based Epidemic Intelligence system* which provides public health event summarization and search capabilities. The data were collected for a four month period, from September 1 - December 31, 2009, by crawling the website. In total 1,397 documents were collected. The data were processed by stripping all boilerplate and markup code using the method introduced in [9].

3.2 Feature Set

In the experiments, the algorithm is run on a feature set consisting of named entities. The entities have been extracted using two different named entity recognition tools: UMLS MetaMap[1] and OpenCalais[2]. More specifically, the features used for our experiments were the *medical condition* and all the variants of the *location* from OpenCalais. MetaMap was used to identify the features *victims*.

3.3 Objectives and Experimental Results

In the following sections, we describe the experiments performed on the introduced dataset and we present the results obtained.

3.3.1 Experiment I: Selection of k

A key consideration in a retrospective event detection is the determination of the number of events k to use as input for the generative model. The choice of the number of events can affect the interpretability of the results. For example, a solution with too few events will generally result in very broad events. A solution with too many events will result in un-interpretable events that pick out idiosyncratic feature combinations. We used a hill-climbing approach to discover the number of events. This method detects all peaks on the articles count-time distribution, and then computes salient scores for each of them. A proportion of the number of salient peaks is then used as an initial estimation of the number of events; such a proportion is an experimentally determined parameter. The objective of this experiment is to determine the value of k for which the algorithm performs best.

We ran our method ten times, each time setting a different input value for the number of events k. Then, we had the best response from our approach with $k = 15$, which correspond to the 50% of all peaks on the articles count-time distribution. Based on this result, we chose $k = 15$ for the manual assessment.

3.3.2 Experiment II: Cluster Quality

A ground truth data set for public health event detection is unavailable. Therefore, to evaluate the correctness of cluster-document assignment, we performed a manual evaluation. The algorithm was applied to our data set and fifteen clusters were created. These clusters were manually assessed by three subjects who had to decide, for each document, whether it was assigned to the correct cluster. In more detail, for each created cluster, the testers were confronted with the two most probable entities for disease, location and victim. These six terms were considered to be descriptive for the cluster, or the documents belonging to this cluster, respectively. The testers had to decide whether the document under consideration described an event taking place in

[1] `http://mmtx.nlm.nih.gov/`
[2] `http://www.opencalais.com`

the location specified by the cluster term labels. They had to decide whether it dealt with the disease and the victims mentioned in the cluster description. When all three criteria were fulfilled, the tester was asked to label this document as correctly assigned. The quality of cluster assignment was measured in terms of precision and recall.

The manual evaluation of the fifteen clusters resulted in an average precision of 65.3% and an average recall of 80.7%. The values differ between the single cluster: for three clusters, precision values of 90% or more were achieved. The best precision and recall values (91.6% and 95.7%) were achieved for the cluster characterized by the terms *china, beijing, flu, swine flu, people, female* which clearly reflects to swine flu cases in China. The lowest precision of 11%, but highest recall (100%) values were achieved for the cluster with the terms *japan, tokyo, disease, cholera, people, children*. The documents assigned to this cluster reported cholera outbreaks, but in other locations than Japan.

This shows that the two most probable terms per feature type selected to describe the content of the cluster based on their probability are not always reflecting the content correctly. A reason might be that some articles are represented by a very large set of features. Therefore, it is difficult to automatically select for the features representing the content best.

4. RELATED WORK

4.1 Event-based epidemic intelligence

Numerous systems exist to detect public health events, notably, PULS [6], Proteus-BIO [5], and BioCaster [4]. However, none of these approaches consider the use of an unsupervised event detection approach. In contrast to them, we propose an approach that exploits unsupervised machine learning technology. Thereby, allowing $PHEs$ to be identified even if no matching keyword or linguistic pattern can be found.

4.2 Unsupervised event-detection

In text mining, the problem of event detection has been examined using news articles as part of a broader initiative named Topic Detection and Tracking [1]. The holy grail in this body of work, has been to automatically acquire a landscape view of a document collection, which answers, in a compact manner, the questions of: "What Happened? and "What is New?".

The event detection task can be divided into two categories: retrospective and new event detection, in either online or off-line mode [13]. Retrospective refers to the detection of previously unidentified events from an accumulated historical collection. New event detection refers to the discovery of the onset of new events, either from live feeds in real-time (online model) or under a closed-world assumption.

Two main approaches have been considered to solve the problem of event detection, namely: feature-based [10, 7] or document-based [1, 13]. In document-based approaches, event detection events is done by clustering documents based on semantics and time stamps. In feature-based approaches the temporal and document distributions of words are first studied and events of words are discovered using either the distribution of the feature over the time, namely trajectory, or a generative model of the features [10]. In order to address the requirements of Epidemic Intelligence, we incorporate aspects from both the trajectory and generative model approaches.

5. CONCLUSIONS

In this work, we draw attention to some of the problems faced in the area of Event-Based Epidemic Intelligence. Particularly, we shed a new light on the task of *Public Health Event Detection*

In order to overcome the problems faced with current Event-Based Epidemic Intelligence systems, we propose a hybrid unsupervised event detection, which combines aspects from two different feature-based detection approaches.

We evaluated our approach: according to the quality of article clusters, our results show that we are able to achieve a precision of 66% and a recall 81% on manually annotated data. Qualitative assessment of the events also show that, in fact, they correspond to real-world health events, that have been listed on the public bulletin of international agencies.

6. REFERENCES

[1] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR*, pages 37–45, 1998.

[2] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the royal statistical society*, 39(1):1–38, 1977.

[3] D. H. et al. The landscape of international event-based biosurveillance. *Emerging Health Threats*, 2009.

[4] N. C. et al. Biocaster: detecting public health rumors with a web-based text mining system, 2008.

[5] R. G. et al. Information extraction for enhanced access to disease outbreak reports. *J. of Biomedical Informatics*, 35(4):236–246, 2002.

[6] R. S. et al. Text mining from the web for medical intelligence. *Mining Massive Data Sets for Security*, 19:295–310, 2008.

[7] Q. He, K. Chang, and E.-P. Lim. Analyzing feature trajectories for event detection. In *SIGIR*, pages 207–214, 2007.

[8] T. Hofmann. Probabilistic latent semantic analysis. In *UAI*, pages 289–296, 1999.

[9] C. Kohlschütter, P. Fankhauser, and W. Nejdl. Boilerplate detection using shallow text features. In *WSDM*, pages 441–450. ACM, 2010.

[10] Z. Li, B. Wang, M. Li, and W.-Y. Ma. A probabilistic model for retrospective news event detection. In *SIGIR*, pages 106–113, 2005.

[11] C. Paquet, D. Coulombier, R. Kaiser, and M. Ciotti. Epidemic intelligence: a new framework for strengthening disease surveillance in Europe. *Euro Surveillence*, 11(12):212–214, 2006.

[12] M. Steyvers and T. Griffiths. *Probabilistic Topic Models*. Lawrence Erlbaum Associates, 2007.

[13] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *SIGIR*, pages 28–36, New York, NY, USA, 1998. ACM.

[14] D. Zhang, C. Zhai, J. Han, A. Srivastava, and N. Oza. Topic modeling for olap on multidimensional text databases: topic cube and its applications. *Stat. Anal. Data Min.*, 2(5-6):378–395, 2009.