

Unsupervised Public Health Event Detection for Epidemic Intelligence

M. Fisichella, A. Stewart, K. Denecke, W. Nejdl

Abstract

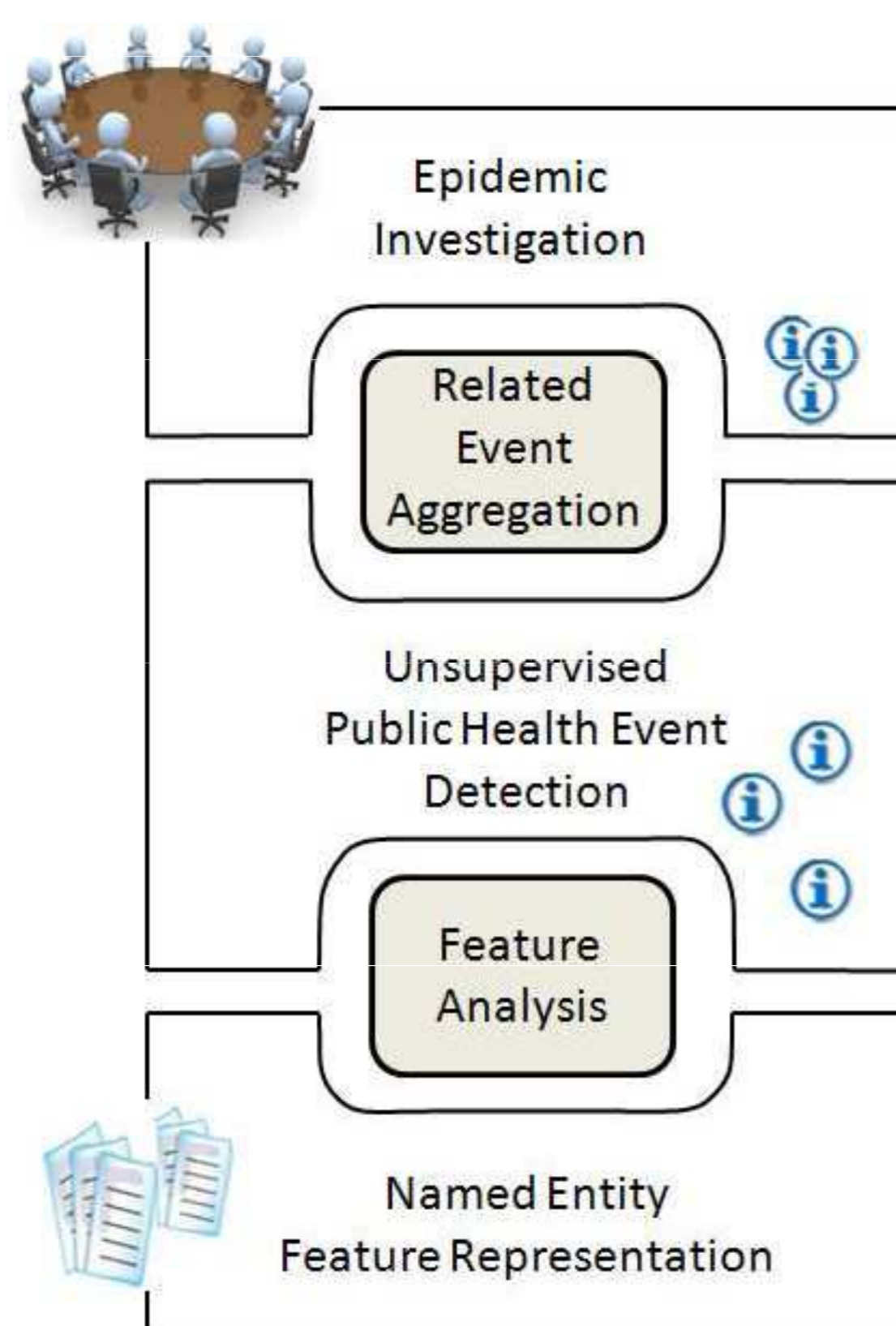
Recent pandemics such as Swine Flu have caused concern for public health officials. Given the ever increasing pace at which infectious diseases can spread globally, officials must be prepared to react sooner and with greater epidemic intelligence gathering capabilities. However, state-of-the-art systems 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.

Unsupervised Public Health Event Detection Process

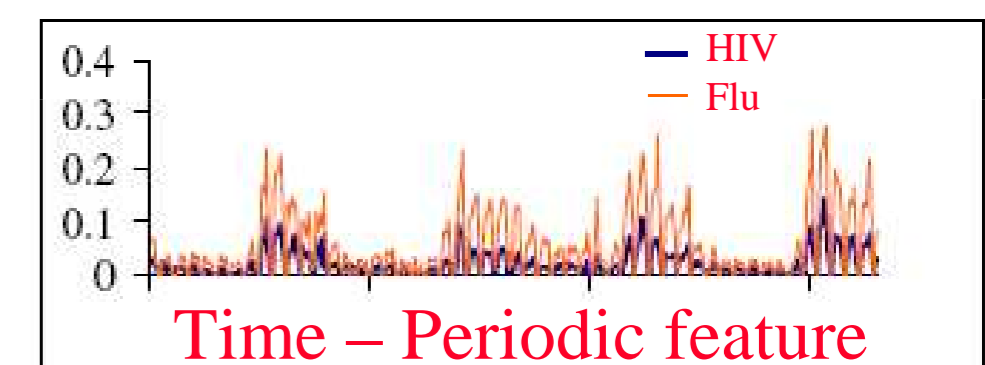
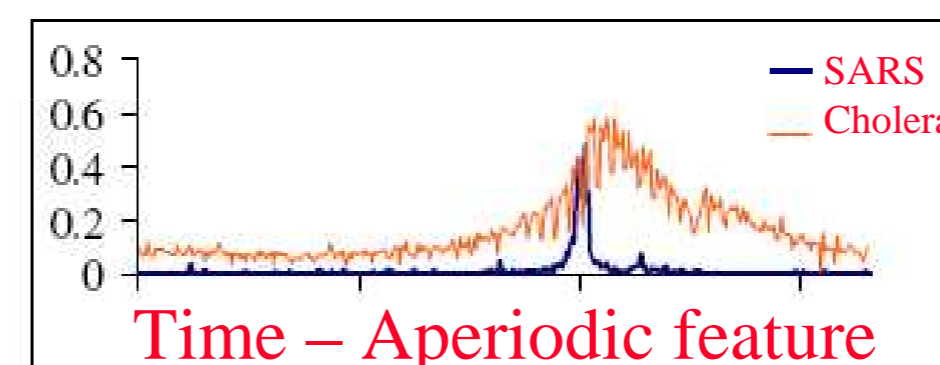
We consider a three stage process for detecting unsupervised public health events. In the first stage, **Named Entity Feature Representation**, we build **entity-centric document surrogates** that are suitable for the **medical domain**.

The manner in which we extract features and represent documents is outlined in next section. We then perform **Feature Analysis** on the extracted set of features to prune the less relevant ones. 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.

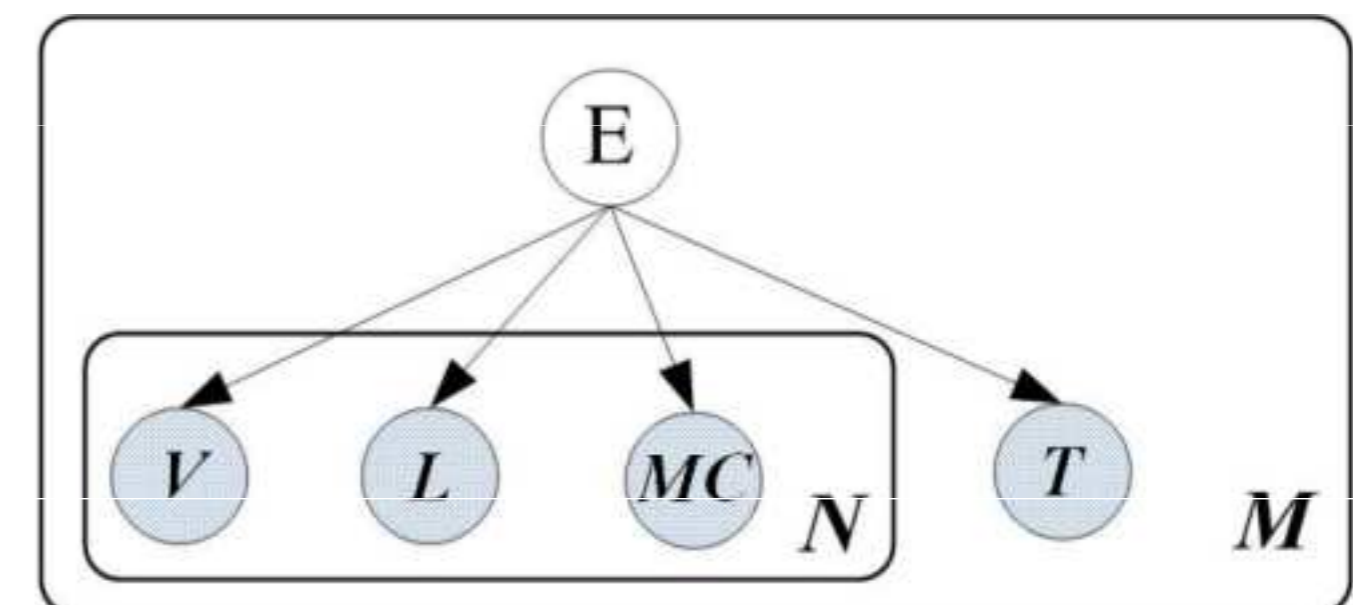


1- Named Entity Feature Representation

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, A as well as a Health Event Template, T .



The template 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 (medical conditions), where (locations) and when (time, defined as the period between the first relevant article and the last relevant one).



2 - Feature Analysis

Spectral analysis is a common technique for identifying periodic and aperiodic features. In this approach features are classified with respect to their periodicity (Pw) and their dominant power spectrum (Sw). For each **aperiodic feature**, we keep only the bursty period, which is modeled by a **Gaussian Distribution**, while for the **periodic features** we chose a mixture of K **Gaussian** distributions, where $K = \lfloor P/Pw \rfloor$.

3 - Detecting Public Health Events

A core step, in the unsupervised detection of events is the clustering of the articles and generation of events using a **Generative Model for Health Events**.

Algorithm 1: The generative model for unsupervised Event Detection

```

begin
  Choose an event  $e_j \sim \text{Multinomial}(\theta_j)$ ;
  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  $victim_{i,v} \sim \text{Multinomial}(\theta_p|time_i)$ ;
    Choose a  $disease_{i,d} \sim \text{Multinomial}(\theta_d|time_i)$ ;
    Choose a  $location_{i,l} \sim \text{Multinomial}(\theta_l|time_i)$ ;
  end

```

The intuition to inject into the generative model the bursty period of the features allow us to compute an initial starting point that, estimated in this way, can be expected to be closer to the optimum (of the well known **Expectation Maximization** algorithm) than a randomly picked initial point.

Acknowledgement: This work was funded, in part, by the European Commission Seventh Framework Programme (FP7/2007-2013) under grant agreement No.247829.