Detecting Public Health Indicators from the Web for Epidemic Intelligence

Avaré Stewart, Marco Fisichella, and Kerstin Denecke

L3S Research Center, Appelstr. 9A, Hannover, Germany
{stewart,fisichella,denecke}@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 an approach that shifts the paradigm for how public health events are detected. Instead of manually enumerating linguistic patterns to detect public health events in human language text (pattern matching); we propose the use of a statistical approaches, which instead learn the patterns of public health events in an automatic or unsupervised way.

Keywords: Epidemic Intelligence, Surveillance and Analysis.

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. Emerging infectious diseases are those considered to be either: completely new, resistant, or reoccurring. Only an early detection of disease activity, followed by a rapid response, can mitigate the impact of epidemic threats [1]. As a result, the multi-disciplinary area of Event-Based Epidemic Intelligence (EI) has emerged as a body of work devoted to the early identification of potential health threats from unstructured text that is present on the Web.

State-of-the-art systems in EI are *Automatic Event-Based* systems [1]. The algorithms used in these systems typically detected disease related activity, by relying upon predefined templates, such a keywords or patterns, within the text. However, a major drawback of this is that the only indicators about public health events a system is capable of detecting, are those that are explicitly under surveillance. This limitation poses a problem, for example, for an early detection system if a disease is emerging and can only be characterized by symptoms and has no known name. The first steps toward overcoming this limitation is to view the Epidemic Intelligence in a new light.

M. Szomszor and P. Kostkova (Eds.): E-Health 2010, LNICST 69, pp. 10–17, 2011.

1.1 Proposed Solution

In this work, we address this challenge, by seeking to learn patterns in an automatic and unsupervised way, which can then be used as indicators for the presence of a public health event. Instead of using keywords, we use an entity-centric unsupervised learner to automatically detect salient patterns within a document collection. These patterns are intended to be an indication that disease-related activity is occurring; thus, we refer to this task as ***Public Health Indicator Detection***. More specifically, we address the following questions:

1. *How can we characterize a Public Health Indicator*
2. *How do we measure the quality of Public Health Indicators?*
3. *When is one set of indicators better than another?*

2 Discovering Public Health Indicators

In Figure 1, an overview of Public Health Indicator Detection is depicted and outlined in Algorithm 1. Each stage of our algorithm is discussed in detail below.

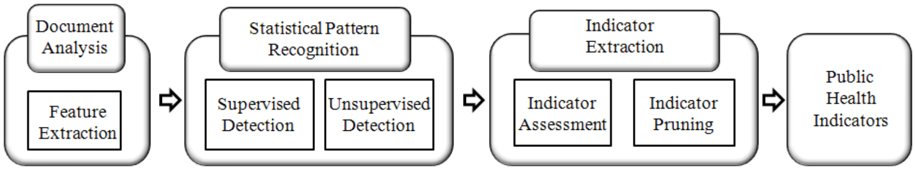

Fig. 1. Overview: Generating Public Health Indicators

2.1 Document Analysis

Given a finite set of text articles, $\mathcal{A}$, we process the raw text of each to build a vocabulary, $V_{\mathcal{T}}$, of relevant terms. The relevance of a term is determined by a set of allowable types, $\mathcal{T}$, which we refer to as a **HealthEventTemplate**. The types considered are: *Location*, *Medical Condition* and *Victim.*

Using the type system and vocabulary, each article is transformed into a vector format. Each entry in the vector corresponds to the frequency with which an entity, of the given type, appears in the article. The document surrogates for the set of articles, $D_{\mathcal{A}} := [|\mathcal{A}|][|\mathcal{T}|][|V_{\mathcal{T}}|]$ are then finally created from the frequency vectors.

2.2 Statistical Pattern Recognition

A key stage in our approach is the use of statistical pattern recognition to discover events. We define an event to be a pattern of entities, which co-occur with

Algorithm 1. Public Health Indicator Generation

```
Input: HealthEventTemplate: $\mathcal{T} := \{Location, MedicalCondition, Victim\}$
Collection of articles: $\mathcal{A} := \{\boldsymbol{a_1}, \cdots \boldsymbol{a_n}\}$, where each $\boldsymbol{a_i} := \{e_1 \cdots e_m\}$, where
each $e_i$ is an entity, given by a type in $\mathcal{T}$
K: desired number of candidate indicators
QualityMeasurementPairs: $\mathcal{F} := \{(f_1, \alpha_1)...(f_n, \alpha_n)\}$
Pattern Recognition Engine: $\varPhi(D_{\mathcal{A}}, K) \models \mathcal{I}^K$
Output: $\mathcal{I}^P$, Set of Public Health Indicators
begin
    // Feature Extraction:
    Hashtable: $V_T := \{\langle$ key=t, $value_t = \{e_1 \cdots e_m\}\rangle\} = \emptyset$
    for each $\boldsymbol{a} \in \mathcal{A}$ do
        for each $t \in \mathcal{T}$ do
            $\mathcal{W}_{\boldsymbol{a},t} = \{e_i | e_i \in \boldsymbol{a} \wedge type(e_i) = t\}$
            $V_T.put(t, V_T.get(t) \cup W_{a,t})$
    $D_{\mathcal{A}} := [|\mathcal{A}|][|\mathcal{T}|][|V_T|]$,construct document surrogates
    // Pattern Recognition:
    $\varPhi(D_{\mathcal{A}}, K) \models \mathcal{I}^K$
    // Indicator Assessement:
    for each $I \in \mathcal{I}^K$ do
        for each $(f_i, \alpha_i) \in \mathcal{F}$ do
            if $Quality_f(I) \geq \alpha$ then
                $\mathcal{I}^P = \mathcal{I}^P \cup \{I\}$
end
```

such saliency, that an unlabeled, real-world event, can be inferred from the content of the articles that contain mentions of these entities. Since the set of documents that describe the same event contain similar sets of term co-occurrences, the documents themselves cluster.

We propose that these patterns can be found in a statistical manner for public health, without the need for defining linguistic templates to extract the co-occurrence of entities from the text. The statistical patterns we find, (i.e., clustering of documents) is considered to be a "hint" that a potential public health event has occurred, or is currently occurring.

As denoted in Figure 1, pattern recognition may be accomplished using either a supervised or unsupervised approach [2]. In general the process of recognizing these patterns is taken to be a mapping of each document surrogate to one or more of the K different clusters in the IndicatorCandidate set, $\mathcal{I}^K$. We now present the following definitions for an IndicatorCandidate, Indicator, and PublicHealthIndicator, as follows:

Definition 1. *Let an* ***IndicatorCandidate*** *set,* $\mathcal{I}^K := \langle \boldsymbol{C}, \boldsymbol{D}, \varPhi \rangle$*, be a set derived from a pattern recognition engine,* $\varPhi$*: where* $\boldsymbol{C}$ *is a set of K clusters;* $\boldsymbol{D}$ *is a set of documents surrogates;* $\varPhi : \boldsymbol{D} \xrightarrow{w} \boldsymbol{C}$ *is a mapping of the document* *surrogates to one or more of the clusters, $C = \{c_1 \cdots c_K\}$; w is weight representing the confidence associate with the assignment of the surrogate to a cluster. In general, we say an* ***Indicator****, I, is a subset of the IndicatorCandidate set, such that the* $|I| = 1$.

Based on Definition 1, potentially many Indicators are produced in the IndicatorCandidate set. **PublicHealthIndicators** is a the subset of IndicatorCandidates, which are filtered according to some criteria for their goodness or quality.

2.3 Indicator Extraction

When is an Indicator good? This question is particularly important when the statistical patterns are recognized in an unsupervised manner since, in general, many clusters may be produced - even if there are no natural patterns in the data.

Since all indicators may not have the same quality, we define Indicator Extraction as the two stage process of: 1) defining a quality measure to apply to IndicatorCandidates (Indicator Assessment) and 2) selecting a subset of the IndicatorCandidates as *PublicHealthIndicators* (Indicator Pruning).

Quantitative Assessment. We assess the quality of Indicators based on two criteria: quantitative and qualitative. Quantitatively, the quality of an Indicator can be determined, given a set of **QualityMeasurementPairs** $(f_i, \alpha_i) \in \mathcal{F}$, where each f_i is a measurement and α is a threshold value for interpreting when the measurement is of a good quality. Based on the application of such a measure to one or more Indicators, we prune the IndicatorCandidates to generate a PublicHealthIndicator according to the following:

$$PublicHealthIndicator = Quality_f(I) \geq \alpha \tag{1}$$

A number of measures can be used, to determine the quality. For example, precision and recall can be used to assess the quality of the generated indicators, the Response set (Res), with respect to an alternative clustering of the articles, the Reference set (Ref) [3] as followings:

$$Recall(Ref, Res) = \frac{\sum_{c_i \in C_{Ref}} |c_i| - overlap(c_i, Res)}{\sum_{c_i \in C_{Ref}} |c_i| - 1} \tag{2}$$

Qualitative Assessment. Recall from Definition 1, that an optional weight, w, may be used to associate a document to a cluster. In the qualitative assessment of Indicators the statistics describing the distributions of these weights and their overall magnitude are taken into account. We express this in terms of 1) the sum of the weights in a given interval and 2) entropy. Entropy is the measure of the uncertainty or the amount of disorder associated with a distribution. Specifically, a high entropy value means that the articles associated to the given indicator cluster have diverse probabilities and, if we consider that the samples are sorted in a descending order, this describes a distribution which decreases rapidly. On the other hand, a low value denotes similar probabilities of articles associated to the given indicator. Entropy is defined as follows:

$$H = - \sum_{i=1}^{N} w \log_2 w \tag{3}$$

3 Experiments

The goal of our experiments is twofold: first to qualitatively compare the indicators extracted with an unsupervised approach to a state-of-the-art, template-based extractions. Second, we qualitatively, assess how the magnitude of the weights associating a document to a cluster influence the quality of an indicator.

3.1 Experimental Setting

Data Sets. To build our document collection, we downloaded the web pages for each urls listed in source column of the PULS fact base [4], for the period from, January 1 - December 31, 2009. Of the 2,587 documents collected, we used the 1,280 documents for which the PULS date column could be automatically computed as a timestamp. For the same time period, we also collected the records present in the PULS fact base, to use as a benchmark. We used both the OpenCalais and UMLS MetaMap entity extraction tools. Since MetaMap produces multiple named entity candidates, a deterministic choice for selecting the correct annotation automatically is error prone. On the other hand, we found that OpenCalais does not recognize entities as victims, but has a high precision for detecting the other entities given by the $HealthEventTemplate$.

3.2 Template Matching Benchmark

The benchmark system aggregates facts into the same group, or equivalence class, if they share the same disease and county, within a temporal window of 15 days. Based on this criteria, the records of the PULS fact base that we collected, yielded a total of 524 clusters and 3,722 documents. From the 524, we used those clusters that constrained at least 10 documents; this amounted to 70 clusters.

3.3 Statistical Pattern Recognition

Numerous techniques exist for detecting events in an unsupervised way. In this work, we base our unsupervised event detection algorithm on the Retrospective Event Detection [5] algorithm. This model for event detection, provides a framework for handling the multiple entity types. We extend this method to handle those defined by the $HealthEventTemplate$.

3.4 Results

Part I: Quantitative Assessment of Indicators. The goal of this experiment was to compare the quality of the indicators that were discovered with our approach, to those that were extracted using a template-based method. Using PULS as the Reference and our indicators as the Response, we computed according to Equation 2.

Figure 2 shows the precision and recall for various clusters sizes.

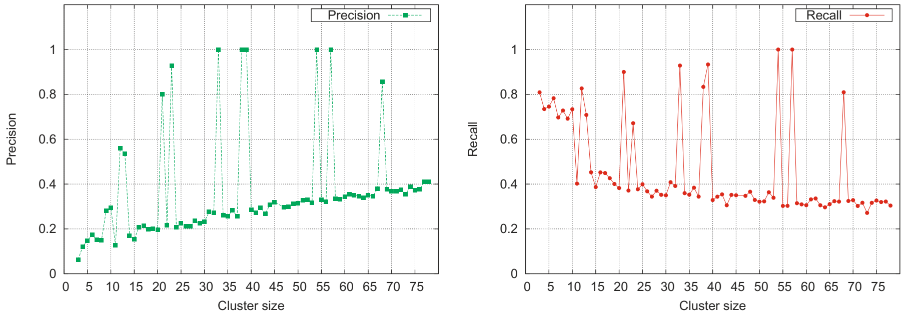

Fig. 2. Extrinsic Indicator Evaluation wrt. Template-based Detection

The first observation we make is that the overall precision and recall seem fairly low for most of the clusters, the majority of the values falling in the range of .02 to .04. This can be explained by the fact that the benchmark set used only two entity types (disease and location) to cluster indicators, where we used three entity types, as defined by the $HealthEventTemplate$. This would suggest that further experiments are needed to select the same entity types as in the benchmark set.

Also, we notice that there are several spikes in the graph for both precision and recall reaching a maximum value of 1, for different values of K. Upon closer inspection, we notice that when the precision or recall reached these values the contribution of the *victim* entity type had a much smaller contribution out of the three that we used in our $HealthEventTemplate$. We also notice that there are several values of K for which an alignment above .8 occurs. We believe this already shows promising results that the statistical approach does,at least, align with a template-based approach.

Part II: Qualitative Assessment of Indicators. In Figure 3a, the weights are expressed by the probability of a document, given and event, according to Retrospect Event Detection algorithm mentioned in Section 3.3. In Figure 3 we show the results for each quartile, given a randomly selected event.

As can be seen, the larger magnitude weights are mainly contained in the first quartile, while the values for the other quartiles, related to different values of K, are almost zero. For a small value of K, we notice bigger values for the sum

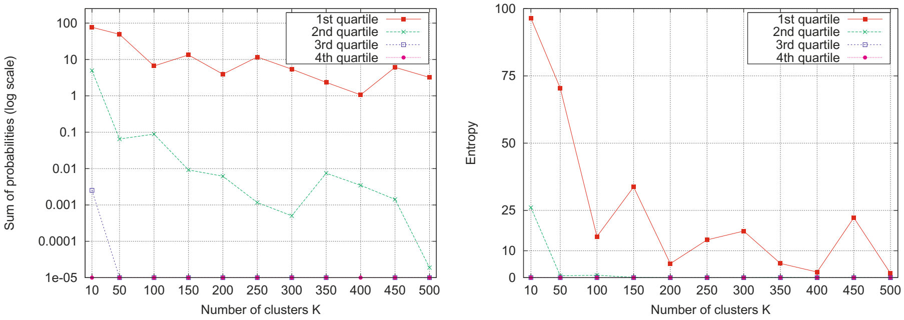

Fig. 3. Qualitative values over the number of clusters K

of probabilities. This is due to the fact that for small values of K, each cluster is represented by a larger grouping of documents - hence more probabilities are summed. On the surface, this would suggest that a smaller value of K is better, however, this value does not reveal any information about the order (or disorder) among the probabilities of documents associated with the event. To examine this, we compute the entropy (Figure 3b). Such a measure indicates the disorder for each quartile, zero entropy being the best value. As can be observed from this figure, a small value of K can have a high entropy value; while for $K = 500$, the entropy is almost zero.

3.5 Discussion

These preliminary results suggest that an unsupervised approach to detecting public health indicators can, at least, align with indicators that have been detected with a template base approach. Also, we say that based on a extrinsic qualitative evaluation, we would prune indicators that have a precision and recall below 80%.

In the unsupervised approach, we produce many more indicators than in the template approach given the number entities defined in the $HealthEventTemplate$. Further evaluation is need for the non-overlapping indicators we detected to evaluated. Finally we note that numerous systems exist to detect public health events [1,6]. None of these existing Event-Base EI systems use an unsupervised event detection approach. As such, they do not allow for public health events to be identified in the absence of predefined matching keywords or linguistic rules.

4 Conclusions and Future Work

We introduce our approach to the discovery of public health indicators; and presented formalizations for characterizing public health indicators. We realized the approach by discovering indicators in an unsupervised manner and further present a framework to evaluate the their quality. We have shown that a statistical approach to detecting public health indicators can produce indicators that are similar to a template matching algorithm. The impact of this work is that epidemic investigators can now rely upon alternative sources and techniques to corroborate information about public health events. This is important, since a diversity of information sources and detection techniques can offer an additional means of mitigating the impact of potential threats.

In future work, a more detailed evaluation of the proposed algorithm will be undertaken. This includes additional measures, such as the B-Cube for computing precision and recall. Also, it should be noted that many factors influence the quality of public health indicators. For example, the existing prevalence levels of a disease or even the personal preference, of the information seeker: such as their geographical location or occupation. Assessing the quality of an indicator based on such factors requires a more robust qualitative evaluation with input from domain experts. We plan this as future work.

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

References

1. Hartley, D., Nelson, N., Walters, R., Arthur, R., Yangarber, R., Madoff, L., Linge, J., Mawudeku, A., Collier, N., Brownstein, J., Thinus, G., Lightfoot, N.: The landscape of international event-based biosurveillance. Emerging Health Threats (2009)
2. Jain, A.K., Duin, R.P.W., Mao, J.: Statistical pattern recognition: A review. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(1), 4–37 (2000)
3. Bagga, A., Baldwin, B.: Algorithms for scoring coreference chains. In: Language Resources and Evaluation Workshop on Linguistics Coreference, pp. 563–566 (1998)
4. Yangarber, R., von Etter, P., Steinberger, R.: Content collection and analysis in the domain of epidemiology. In: International Workshop on Describing Medical Web Resources (2008)
5. Li, Z., Wang, B., Li, M., Ma, W.-Y.: A probabilistic model for retrospective news event detection. In: SIGIR 2005, pp. 106–113 (2005)
6. Linge, J.P., Steinberger, R., Weber, T.P., Yangarber, R., van der Goot, E., Khudhairy, D.H.A., Stilianakis, N.I.: Internet surveillance systems for early alerting of health threats. Eurosurveillance 14(13) (2009)