Making Use of Social Media Data in Public Health

Kerstin Denecke
L3S Research Center
Appelstrasse 9a
Hannover, Germany
denecke@L3S.de

Peter Dolog
Aalborg University
Selma Lagerlöfs Vej 300
Aalborg, Denmark
dolog@cs.aau.dk

Pavel Smrz
Brno University of Technology
Bozetechova 2
Brno, Czech Republic
smrz@fit.vutbr.cz

ABSTRACT

Disease surveillance systems exist to offer an easily accessible "epidemiological snapshot" on up-to-date summary statistics for numerous infectious diseases. However, these indicator-based systems represent only part of the solution. Experiences show that they fail when confronted with agents that are new emerging like the agents causing the lung disease SARS in 2002. Further, due to slow reporting mechanisms, the time until health threats become visible to public health officials can be long. The M-Eco project provides an event-based approach to the early detection of emerging health threats. The developed technologies exploit content from social media and multimedia data as input and analyze it by sophisticated event-detection techniques to identify potential threats. Alerts for public health threats are provided to the user in a personalized way.

Categories and Subject Descriptors

J3 [**Life and Medical Sciences**, H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Epidemic Intelligence, Social Web, Webscience

1. BACKGROUND

Many factors in today's changing societies contribute towards the continuous emergence of infectious diseases. In 2009, in the Federal Republic of Germany alone, approximately 400,000 cases of infectious diseases have been reported; 420 public health departments were engaged in collecting the corresponding data and processing the reports. Systems such as SurfStat [1], provided by the Robert-Koch-Institute, the national health organization in Germany, exist to offer an easily accessible "epidemiological snapshot" on up-to-date summary statistics for numerous infectious diseases. However, these indicator-based systems represent only part of the solution. Experiences show that they fail when confronted with agents that are new emerging like the agents causing the lung disease SARS in 2002. Further, due to slow reporting mechanisms, the time until health threats become visible to public health officials can be long.

In response, Epidemic Intelligence (EI) has emerged as a type of intelligence gathering which aims to detect events of interest to the public health, from the unstructured text of news and outbreak reports. More specifically, event-based surveillance systems use additional sources of information and are intended to improve the early detection of potential health threats. text. The events which are considered to be relevant for detecting an emerging disease are annotated with additional information (such as threat or severity level) and then aggregated to produce signals. The signals are intended to be an early warning against potential public health threats, and the epidemiologist uses them to assess a risk; or corroborate and verify the information locally and with international agencies.

In such a scenario, the diversity of the sources plays an important role in the intelligence gathering process. Online news are already under monitoring by disease surveillance systems such as MedISys [2]. Other present-day systems use news and outbreak reports as sources of information to support intelligence gathering (GPHIN, ProMED-Mail [3]).

Medicine 2.0, social medical blogs and other forms of user generated content can be seen as an additional source – but remained so far unconsidered in surveillance systems. However, these sources are of significance, since those who experience as well as treat disease first hand, describe their experiences in blogs and other forms of social media. Thus, this information could provide early indications of potential health threats or additional information. The M-Eco project is intended to provide technologies that enable exploiting social and multimedia data for disease surveillance purposes.

There are several open issues and challenges to address when considering this data for disease surveillance. Huge amounts of data need to be processed; relevant pieces of information need to be carefully selected; natural language needs to be interpreted even if it is common language. These challenges are addressed within M-Eco.

2. M-Eco INNOVATIONS

M-Eco targets at complementing traditional surveillance systems with additional approaches for the early detection of emerging threats. It addresses limitations of current systems for Epidemic Intelligence by

- Exploiting more sophisticated event-detection technologies (unsupervised and supervised methods),
- Monitoring additional resources (Web 2.0 data, multimedia),
- Enabling access to additional information related to disease outbreaks collected from multiple sources, and
- Personalizing and filtering results.

All technology developed in the project are designed as processing services that could be included in existing surveillance systems. These services are expected to improve epidemic intelligence including intelligent and adaptable interfaces for data output, visualization, navigation and decision support. The technology development focuses on the processing of two languages: German and English and on infectious diseases. Within the project, integration with MediSys [2] and SurfStat [1]

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
WWW 2012, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

will be realized. Further, the M-Eco Portal will integrate all services developed.

The assumption behind the project is that through social and multimedia channels information on potential health threats can be detected earlier. This hypothesis will be studied. It will be assessed to what extend social media data can contribute to the early detection of public health threats.

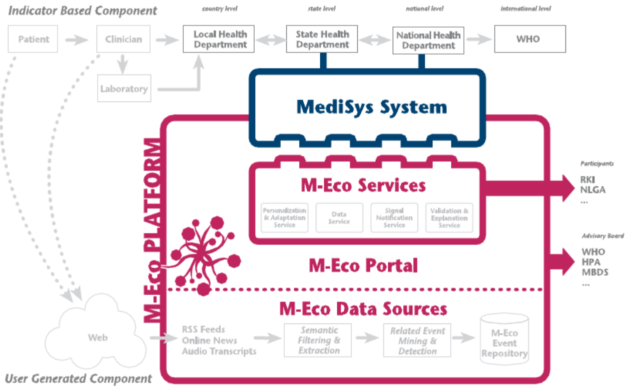


Figure 1: M-Eco services and interactions

3. SCENARIO AND USERS

The users of the M-Eco system will be epidemiologists, public health officials or decision makers. Representatives of health organizations are supporting M-Eco in its requirement analysis and evaluation (e.g. Robert Koch-Institut (Germany), World Health Organization, European Centre for Disease Prevention and Control).

The M-Eco scenario focuses on creating user-defined signal searches. The user has to specify a signal search, i.e. he selects symptoms or disease names together with a location or time span in which he is interested. Given such signal definition, the system shows related alerts.

M-Eco Signal Search Scenario

Dieter is an epidemiologist working at the Robert Koch Institute, the national institute in charge of surveillance of infectious diseases. Recent monitoring of basic indicator-based surveillance with the institute's SurvNet system has brought attention to a potential outbreak of Legionnaires 'disease, a respiratory disease common during winter, near Aachen, Germany. The user now wishes to compare this finding with other possible co-occurring events, and turns to the M-Eco platform for support.

He is creating a M-Eco signal search, specifying diseases, symptoms of his monitoring interest together with a location or time span in which he is interested. More specifically, he enters for diseases the keywords: "legionellosis," "respiratory," "pneumonia.". Under symptoms, he specifies: "bloody sputum," "ILI," "runny nose," "sore throat" and "cough.". He restricts the time period to "December-March" and the location to "North Rhine-Westphalia (NRW)" and "Netherlands." He also enters his requested sources as being media articles, and sets the alert level to high, but is also interested in sources from blogs and Twitter and marks these as having a low alert level.

As a daily routine the M-Eco system checks its indexed blogs, audio transcripts and other data tracking and detecting topics, locations, and diseases that have been extracted. The M-Eco System discovers that for certain term combinations related to legionnaires' disease this threshold is crossed and generates a signal which is distributed to all subscribers to the respective channel. Signals and related documents are presented in an intuitive and accessible way: tag clouds provide a quick overview of the occurrence of matching events and frequency. An event summary template provides an overview of relevant information.

4. M-Eco ARCHITECTURE

The M-Eco system comprises four major components that realize the individual processing steps (Fig. 1):

- Content collection and preprocessing,
- Event detection,
- Signal generation,
- Recommendation and user modeling.

The components interact via web services. Collected (textual) content and processing results are stored in a database. Figure 2 shows the information flow between the single components. For simplification reasons, the database accesses are not shown in the diagram. Knowing the user interest specified in a signal definition provided by a user, the system continuously monitors the incoming text and data streams for relevant events. Once patterns of interest are identified, appropriate services for pattern analysis and interpretation are triggered and an alert is produced. The single components will be described in the following.

The *Content collection component* collects continuously data from various sources including TV, radio, online news, blogs and Twitter. TV and radio data is collected via satellite and transcribed by the SAIL Media Mining System[1]. Medical blog data includes MedWorm[2] listed blogs, manually selected blogs and forums collected through corresponding APIs (e.g., Twitter). Online news data collected by the MedISys system is integrated into the M-Eco data collection as well.

The *Document analysis component* filters and pre-processes the collected (textual) data before making it available for the event detection component. Pre-processing includes filtering of irrelevant data, recognition of mentions of disease names and symptoms, locations, time etc. The latter is realized by OpenCalais[3]. The documents are analysed linguistically by Minipar[4]. As a result, a set of documents annotated with named entities and linguistic structures is produced. These tagged documents are indexed through MG4J [5](Managing Gigabytes for Java) which is a free full-text search engine for large document sets and made available via web services.

The *Event detection and signal generation component* exploits the tagged documents to identify patterns of interest and to produce signals. It works in two modes: The *unsupervised event detection* (introduced in [5]) groups documents into clusters by a retrospective event detection algorithm. These clusters are interpreted as signals and exploited by the recommendation component (see below). The *supervised event detection* considers the signal definition entered by the user. Given this information, it first retrieves data from the data repository that is relevant for

[1] http://www.sail-technology.com/products/commercial-products/media-mining-indexer.html

[2] http://www.medworm.com/

[3] http://www.opencalais.com

[4] http://webdocs.cs.ualberta.ca/~lindek/minipar.htm

[5] http://mg4j.dsi.unimi.it

the specified information need. The system then identifies segments (e.g., sentences, paragraphs) in the relevant documents by means of a supervised machine learning algorithm (see [6] for details). This information is then exploited by standard statistical algorithms for biosurveillance (e.g. CUSUM, Farrington [7]) to produce signals as alerts for health officials.

The *Recommendation component* gets as input the document clusters or the calculated signals and either selects those that are of interest for the user according to his profile or ranks the signals and associated texts appropriately. This component requires the user profile that consists of information on a specified signal definition as well as user feedback from previous searches and user interactions. The ranking of signals in the result presentation is adapted to the user interest and irrelevant signals can be filtered out. The produced signals are presented in the user interface.

The *user interface* allows a user to specify his interest in terms of a signal definition. It collects information on disease names or symptoms, locations to be considered by the surveillance system. Further, the generated signals and related information on indicators and the information source are presented to the user (see Fig. 3). Users are enabled to browse through the results. Various visualization methods are applied to present the results in an easy understandable way (e.g., as word clouds, in maps or graphs). Further, the user is enabled to provide feedback to the results in terms of ratings and tags.

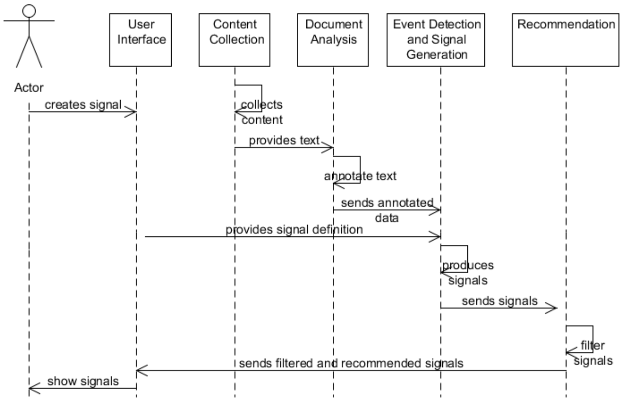


Figure 2: Information flow in the M-Eco System

5. RESULTS

The first prototype of integrated M-Eco components is running, producing signals from online news, Twitter and blogs (see Fig. 3). Single components have been evaluated already. Some of the results are presented in the following. Further, a user assessment studied the relevance of social media data for disease surveillance. Results will be described in this section.

5.1 Relevance Classification

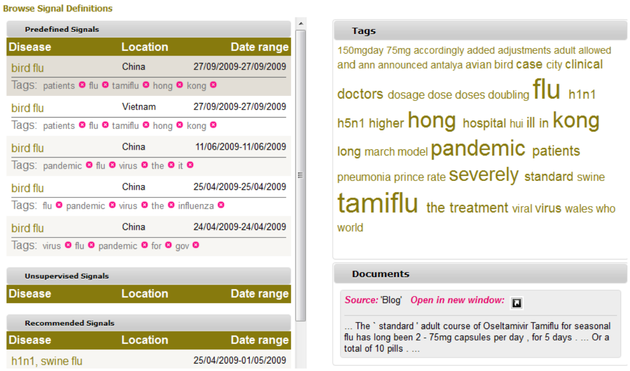
Given the huge amount of information available in the Web, it is a crucial step within the document analysis component to filter out irrelevant texts. For this purpose, a relevance classifier has been developed that exploits machine learning technology. In an experiment the accuracy of relevance classification for tweets has been evaluated. In this experiment, a potentially relevant document is a document mentioning an infection by some disease (or its symptom) stated in a manually created list of diseases and symptoms. The infection can be described from one's own experience or from someone else's point of view (e.g. news feeds on Twitter, documents created by an institution).

This experiment involves the comparison of the classifier against the human annotation (gold standard). 4,000 documents in English and 500 documents in German have been labeled for training the classifier. For testing, 1,000 annotated documents in English and 200 documents in German are available. Precision and recall of the relevance classifier has been calculated. The results of the experiment are stated in the following table 1.

Table 1: Results of relevance classification

	Precision	Recall	F-Measure (F1)
English Tweets	83.3%	74.4%	78.59%
German Tweets	83%	46%	59.2%

As can be seen from the Table 1, the English classifier performs substantially better, which is due to the fact that the English classifier was trained on a larger set of documents. From the experiment also follows that the ratio between relevant and irrelevant documents is about 3:7. Only about 30% of all documents is considered as relevant.



Browse Signal Definitions

Predefined Signals

Disease	Location	Date range
bird flu	China	27/09/2009-27/09/2009
Tags: patients flu tamiflu hong kong		
bird flu	Vietnam	27/09/2009-27/09/2009
Tags: patients flu tamiflu hong kong		
bird flu	China	11/06/2009-11/06/2009
Tags: pandemic flu virus the it		
bird flu	China	25/04/2009-25/04/2009
Tags: flu pandemic virus the influenza		
bird flu	China	24/04/2009-24/04/2009
Tags: virus flu pandemic for gov		

Unsupervised Signals

Disease	Location	Date range

Recommended Signals

Disease	Location	Date range
h1n1, swine flu		25/04/2009-01/05/2009

Tags

150mgday 75mg accordingly added adjustments adult allowed and ann announced antalya avian bird case city clinical doctors dosage dose doses doubling flu h1n1 h5n1 higher hong hospital hui ill in kong long march model pandemic patients pneumonia prince rate severely standard swine tamiflu the treatment viral virus wales who world

Documents

Source: 'Blog' Open in new window:

... The ` standard ' adult course of Oseltamivir Tamiflu for seasonal flu has long been 2 - 75mg capsules per day , for 5 days . ... Or a total of 10 pills . ...

Figure 3: Screenshot of the M-Eco System

5.2 Personalized Recommendation of Signals

Another evaluation was performed to study the usefulness of the recommendation component. The recommendation algorithm infers the user preferences from his set of defined rules [8]. In the evaluation, the participants indicate whether the system-made recommendations match their interest in the H1N1 virus.

The data set mainly consisted of signals related to mentions of the H1N1 virus, but we did add sparse signals on two other diseases, E. Coli and the common flu. In total, we had 138 signals and 419 documents belonging to one of the signals in the test set. In this scenario, we wanted to test, in general, how the recommendation component would perform and, more specifically, whether it would filter out E. Coli, recommending mostly those related to H1N1 for the users who set their signal definitions to this disease and some related to the common flu.

Users were requested to specify a signal definition of their interest which was used to generate personalized recommendations to them. The five participants were asked to evaluate the recommendations according to relevance, correctness and representativeness.

Results showed that the recommendations were relevant for the five users who did set up at least one of their signal definitions to

H1N1, but practically irrelevant to the other three users. This confirmed our expectation, given the dataset focused on H1N1 that the recommendation component would filter out recommendations of other, non-related, diseases. Users that did set up signal definitions to H1N1 also gave high grades to the correctness of the recommended signals and to the representativeness of their respective recommended documents. Table 2 summarizes the results.

The participants considered the recommendations satisfactory but highlighted a number of improvements. Overall, the users would like to have more control over the recommendations received. In that sense, they suggested to provide more information regarding the reasons why the recommendations were generated and more options to assess them.

Table 2: Summary of the results in terms of precision and mean of correctness and representativeness for the ratings given to recommendations matching H1N1 and non-H1N1 signal definitions

Measure	H1N1	Non-H1N1
Precision	0.81	0.27
Correctness	3.25	1.53
Representativeness	3.37	1.33

5.3 Social Media Relevance

One objective of the M-Eco project is to assess the relevance of social media data for the purpose of disease surveillance. Some examples that support this hypothesis of its usefulness have already been identified within the project, for example for the outbreak caused by E.coli that happened in May –June 2011 in Germany.

Another experiment examined a recent community outbreak of Norovirus at a university in Lower Saxony, Germany, to determine if news articles and information posted to the websites Facebook, Twitter and blogs was faster than reporting through established indicator-based surveillance.

We compared information from each day of the outbreak from local and state health authorities to information from Internet sources. News articles were collected from an existing RSS-feed established to collect news on health and infectious disease in Lower Saxony from local online news websites. Related Twitter content was collected by entering outbreak-related search terms into Topsy, a search engine to analyse Twitter content, and Google to search for information from blogs or other Internet forums.

Reporting of suspected cases occurred first to local health authorities before it was featured as information in news and social media sources. These local health authorities do not, however, fall within legal indicator-based surveillance reporting mandates for Norovirus in Germany. So, the first notification in this case occurred faster than established indicator-based surveillance, and news and social media followed.

The initial reports surrounding the canteen's inquiry to the local health department was the first trigger for this topic in social media sources. This effect can be more clearly seen after the official results from the state health department, when much more attention in both news and social media was seen. Thus, as would be expected news and social media tend to be strengthened by information derived from or connected to the authorities.

We conclude, while a faster signal may not be generated, in the absence of indicator-based surveillance, event-based surveillance of news and social media can provide a trigger to a relevant outbreak signal. Social media was also used for information exchange about the event, creating more information and more noise, and potentially amplifying possible signals for monitoring purposes.

6. CURRENT STATE AND CONCLUSIONS

The M-Eco project enhances the capabilities for disease surveillance by technologies that allow monitoring of social and multi media data. First evaluations show the usefulness of social media data for monitoring purposes. A prototype system is running already that could be demonstrated during the WWW 2012 European projects track. The system components are currently improved and evaluated.

7. ACKNOWLEDGEMENT

The M-Eco project (http://www.meco-project.eu) is funded by the EU within the Seventh Framework Programme, Theme ICT-2009.5.2: ICT for Patient Safety. The consortium consists of seven partners: Leibniz Universität Hannover (Germany), Aalborg University (Denmark), Brno University of Technology (Czech Republic), SAIL Labs Technology (Austria), Robert Koch-Institut (Germany), Niedersächsisches Landesgesundheitsamt (Germany), Joint Research Centre (Italy).

8. REFERENCES

[1] Faensen D, Claus H, Benzler J et al.: SurvNet@RKI – a multistate electronic reporting system for communicable diseases. Euro Surveill 2006;11(4):100-3

[2] Steinberger R et al.: Text mining from the web for medical intelligence [Book Section] Mining Massive Data Sets for Security. - Amsterdam : IOS Press, 2008.

[3] Madoff LC: ProMED-mail: an early warning system for emerging diseases. Clin Infect Dis. - 2004. - 2 : Vol. 15.

[4] Collier N, et al.: BioCaster: detecting public health rumors with a Web-based text mining system. Bioinformatics. 2008. - Vol. 24.

[5] Fisichella M, Stewart A, Denecke K, Nejdl W: Unsupervised Public Health Event Detection for Epidemic Intelligence. CIKM'10, October 25-29, 2010, Toronto, Ontario, Canada

[6] Stewart A, Smith M, Nejdl W. A transfer approach to detecting disease reporting events in blog social media. Proc. of the 22nd ACM Conference on Hypertext and Hypermedia, Eindhoven, The Netherlands, June 6-9, 2011

[7] Höhle M: surveillance: An R package for the surveillance of infectious diseases, Computational Statistics (2007), 22(4): 571-82

[8] Lage R, Durao F, Dolog P, Stewart A: Applicability of Recommender Systems to Medical Surveillance Systems. In Proceedings of the second international workshop on Web science and information exchange in the medical web (MedEx '11). ACM, New York, NY, USA, 2011: 1-6