Supporting Epidemic Intelligence, Personalised and Public Health with advanced computational methods

Conference Paper · October 2015

6 authors, including:

Joao Pita Costa
Genialis

Daniela Paolotti
ISI Foundation

Flavio Fuart
Jožef Stefan Institute

Supporting Epidemic Intelligence, Personalised and Public Health with advanced computational methods

Uporaba naprednih računskih metod za podporo osebnega in javnega zdravstvenega varstva ter nadzora izbruhov nalezljivih bolezni

Joao Pita Costa[1], Daniela Paolotti[2], Flavio Fuart[1], Primož Škraba[1], Evgenia Belayeva[1], Inna Novalija[1]

[1] Jožef Stefan Institute, Jamova cesta 39, 1000 Ljubljana, Slovenia. email: joao.pitacosta@ijs.si
[2] ISI Foundation, Via Alassio, 11/C, 10126 Torino, Italy. email: daniela.paolotti@isi.it

POVZETEK

Sodobna tehnologija nam danes dopušča nadzorovanje zdravja s katerim lahko dopolnimo tradicionalne že obstoječe metode zdravstvenega varstva ter javnega zdravstva. V tem članku bomo predstavili nekaj že obstoječih tehnologij, ki temeljijo na topološki analizi podatkov in orodju za spremljanje medijev, nam pa lahko veliko doprinesejo k odkrivanju različnih vrst bolezni. Omenjena tehnologija je bila razvita v sklopu dela Laboratorija za umetno inteligenco Instituta Jožef Stefan in ISI Fundacije v Italiji. Vsa opisana tehnologija je že razvita in se jo lahko neposredno prilagodi za reševanje tako raziskovalnih kot gospodarskih vprašanj na področju zdravstvenih sistemov.

ABSTRACT

Today, our everyday access to technology permits a health monitoring that can complement the traditional methods in Healthcare and Public Health. In this paper, we present some of this available technology, with a particular focus on disease detection, topological data analysis, and media monitoring tools, made available by the AILAB at the JSI and the ISI Foundation. This technology is ready to be adapted to research and commercial problems in the context of health systems.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health, Medical information systems; I.5.1 [**Pattern Recognition**]: Models - *Statistical, Structural;* G.2.3 [**Discrete Mathematics**]: Applications;

General Terms

Algorithms, Design, Theory, Verification

Keywords

epidemic intelligence, epidemiology, topological data analysis

1. INTRODUCTION

The Artificial Intelligence Laboratory [AILAB] has over 40 researchers conducting research in the field of data analysis with an emphasis on text, web and cross-modal data, scalable real-time data analysis, visualization of complex data, semantic technologies, language technologies and sensor networks. In collaboration with the Department of Communication Systems and with the Centre for Knowledge Transfer in Information Technologies, we have established a cross-departmental laboratory for wireless sensor networks [SensorLab]. The goal is to combine technologies for sensor data acquisition, communication between sensor devices, statistical real-time data analysis, semantic technologies, and to enable a wide range of research and development in different application areas, such as energy, ecology, health, transport, security, and logistics. The AILAB has participated in 14 FP6 projects, 28 FP7 projects and in 2015, we have kicked-off four H2020 projects, amounting to 14 on-going EU-funded projects. We coordinated three of those projects in addition to scientific services provided to industry. We have collaborated with various research institutions and industry, including: Stanford University, University College London, Jožef Stefan International Postgraduate School, Quintelligence, Cycorp Europe, LifeNetLive, Modro Oko and Envigence.

The computational epidemiology and public health group at the **ISI Foundation from Turin, Italy [ISI]** has been developing mathematical and computational methods needed to achieve prediction and predictability of disease spreading in complex techno-social systems. Research performed during the FP7 EU project Epiwork has led to the creation of GLEAMviz (http://gleamviz.org), a publicly accessible, large scale, data driven computational framework, endowed with a high level of realism and aimed at epidemic scenario forecast and policy-making. Identified modelling needs have inspired design and implementation of novel data-collection schemes, such as the collection of real-time disease incidence through web applications. The result of this activity is the creation of a network of web-based platforms for participatory influenza surveillance in Europe, Influenzanet.eu (http://influenzanet.eu) developed under the EU FP7 project Epiwork [9]. Data provided by Influenza.net is the main source for real time forecasts of seasonal influenza provided by the Fluoutlook platform (http://fluoutlook.org).

2. VISION

Regular outbreaks of flu pandemics, about three per century, and other epidemic outbreaks of serious public health concern in the 21st century (like SARS 2002-2004, Ebola epidemic in West Africa 2013-, MERS 2012-) have led to an increased need for tools to detect and manage infectious diseases outbreaks.

Online news media analysis tools can detect possible disease outbreaks reported in the news and measure media impact of vaccination campaigns, public awareness rising and other information disseminated by public health authorities. Media monitoring helps the improvement of health surveillance and the detection of potential or emerging threats to public health, including disease outbreaks, early warning signs for an epidemic or rare illnesses. Several successful detections have been reported. For example, early signs of what later turned out to be the outbreak of Escherichia coli in May 2011 in northern Germany were detected by an early-warning system MediSys [6] developed at the Joint Research Centre of the European Commission. The European Centre for Disease Prevention and Control [ECDC] relies on such information in order to timely detect and monitor signals of potential public health threats during mass gathering events.

Today, the health of a population can be assessed in real time through digital traces. The internet has a good picture of the state of health of the population, coming from digital sources, through all of our connected devices, including smartphones [12].

An efficient method for integration of traditional sources (morbidity, epidemiological data) with new, publicly available data and application of advanced IT and mathematical tools is the key to providing useful tools for health professionals. In the BigData era, public health authorities can take profit of the existing technologies and methods for collection, analysis, storage and prediction. Major technology companies already provide health platforms that permit integration of time-stamped and geo-tagged data from different trackers.

Now is the time to pursue the efficient integration of Data Technology and Public Health expertise. A common challenge of all mentioned use-cases is fusion and analysis of "soft", unofficial, data collected either using targeted approaches (such as influenza.net) or by processing massive, publicly available, data sources from which useful information may be extracted (social media, traditional media) and "hard" data coming from accurate sensor measurements or public health sources. Expertise available through the AILAB and its partners could support a strong research and application track on BigData analytics as part of the Slovenian E&M Health strategy.

3. ACHIEVEMENTS

3.1 Data analytics and tools

AILAB has extensive research facilities and equipment for performing research and development in areas of machine learning, data mining, language technologies, semantic technologies and sensor networks. Some examples of software tools for multimodal data analysis developed at AILAB are summarized in Table 1.

Tool	URL	Description
SearchPoint	searchpoint.ijs.si	Clustered visualization of search topics (Figure 1)
XLing	xling.ijs.si	Wikipedia based cross-lingual similarity computation across 100 languages.
Science Atlas	scienceatlas.ijs.si	Web portal exploring the sc. community in Slovenia (Figure 2).
VideoLectures	videolectures.net	The biggest academic online video repository.
NewsFeed	newsfeed.ijs.si	A real-time aggregated stream of semantically enriched news articles
3XL News	ailab.ijs.si/tools/3xl-news/	Global news monitoring and analysis across several languages (iOS app).
EventRegistry	eventregistry.org	A system for real-time collec. and analysis of news published globally.
Twitter Observatory	twitterobservatory.net	A tool for observing, searching, analyzing and presenting social media.
InfluenzaLab	influenzalab.org	A web portal for methods of TDA to Influenza data.

Table 1: AILAB software tools available on-line

The flexibility of the developed technology permits adaptation to specific applications in Healthcare and Public Health as discussed in Sections 3.3 and 3.4.

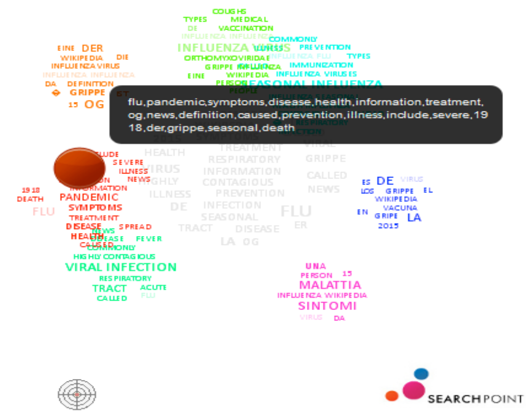

Figure 1 - A screenshot of the portal searchpoint.ijs.si exhibiting the clustered keywords of a query, after searching for the keyword *influenza*.

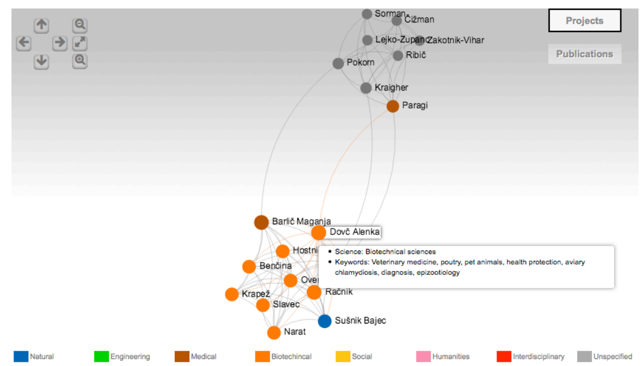

Figure 2 - A screenshot of the portal scienceatlas.si exhibiting the graph of academic collaboration related with the query *influenza*.

3.2 InfluenzaNet

3.2.1 Introduction and European Coverage

An important epidemiological indicator regarding infectious diseases is the number of infected persons over time. The incidence is traditionally based on clinical diagnosis of sentinel medical doctors, but the wide availability of Internet has given rise to new innovative methods. Media attention is generated to encourage people to enroll at one of the national websites by completing an intake questionnaire containing a set of demographic, medical, socio-economic and lifestyle questions. Subsequently, participants receive an email on a weekly basis with a link to a short symptom questionnaire on eventual symptoms since their last visit. When symptoms are reported, some additional questions are presented regarding GP consultation and changes in activities. The incidence for influenza-like illness (ILI) is determined based on the percentage of reporting participants who fit a syndromic case definition. Influenzanet can simultaneously provide ILI incidences based on multiple definitions and in various sub populations, to allow a more nuanced assessment of the ILI activity in the population. Based on weekly symptom reports of tens of thousands of volunteers, the ILI incidence as reported by Influenzanet closely follows the same trends as the ILI incidence as reported by the sentinel network of medical doctors. Based on the intake questionnaire, Influenzanet can determine risk factors for ILI. The risk factors are largely in correspondence with those found in previous studies (high in children and patients with chronic diseases), but also allows the assessment of risk factors which are normally not measured, such as the absence of an increased risk in participants which travel daily by public transport [14]. Based on the vaccination status of every participant, Influenzanet can also provide real-time data on the relative vaccine efficacy in comparison with previous seasons. Since Influenzanet is independent of a GP visiting rate and can apply the same ILI case definition in every country, the ILI incidences can be compared directly between countries. This allows the fitting of transmission models for influenza to be simultaneously fitted to multiple countries and has led to new hypotheses on the relation between influenza, ILI, weather variables, and time of epidemic onset. Integrating multiple data sources could further improve the accuracy of detecting ILI activity. The ILI incidence data from Influenzanet and the sentinel doctors complement each other. The flexibility of the Influenzanet system furthermore allows the collection of data, which goes beyond the monitoring of the ILI activity in the population. All aspects of the system, from monitoring to data analysis and modelling, are readily applicable to other febrile syndromes of infectious origin that afflict other parts of the world, such as dengue fever in tropical regions.

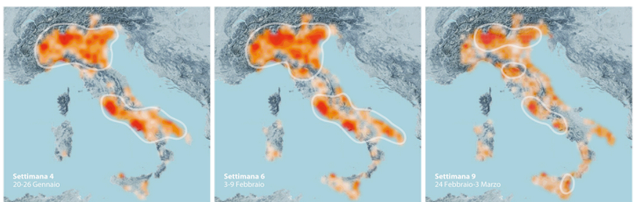

Figure 3 - The website of Influenzanet providing weekly information on Influenza related topics, the real-time information on the evolution of the disease at the local level, and some statistics on the available data.

3.2.2 Analysis

Topological Data Analysis applies qualitative methods of topology to problems of machine learning, data mining and computer vision. In particular, persistent homology is an area of mathematics/computational topology that identifies a global structure by inferring high-dimensional structure from low-dimensional representations and studying properties of an often continuous space by the analysis of a discrete sample of it, assembling discrete points into global structure. When considering a notion of distance on the space, one gets a perspective of the space under different scales, where small features will eventually disappear (and are considered noise). The basic technique encodes topological features of a given point cloud using diagrams representing the lifetime of those topological features (see Figure 4). Recently, these topological methods, explained in detail in [3], have seen a relevant application to the study of the influenza virus as described in [4]. With these methods, we wish to distinguish a trigger before a spike of intensity, or a bifurcation point in the cycle on intensity/widespread of the disease. Each country has a parameterized cycle for the season of flu. In the works presented in [10] and [11], the authors show that TDA provides complementary information to that extracted by the classical quantitative methods used in epidemiology today (see Figure 5).

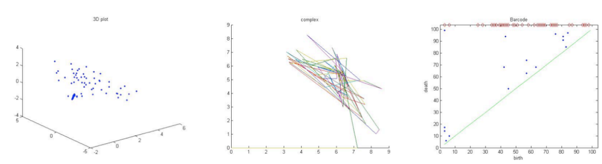

Figure 4 – The pipeline for the computation of topological data analysis for the time series of Italy 2009/10: the given pointcloud of input data (left); the simplicial complex approximating the space of the pointcloud (center); and the correspondent persistence diagram encoding the lifetime of the persistent topological features (right).

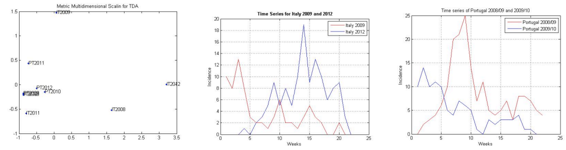

Figure 5. Comparing the flu seasons of Italy and Portugal during 2008-2013 using metric multidimensional scaling (left) to identify: the outlier flu seasons of Italy 2009/10 and 2012/13, with time series plotted for analysis and interpretation (center); the close flu seasons of Portugal 2008/09 and 2009/10 (right).

3.3 Medical News Monitoring

AILAB operates **NewsFeed** [13], an on-line media monitoring system that monitors thousands of web sites and is currently processing around 150.000 news articles daily in many languages. Article text is extracted from web pages and then analysed and annotated in order to produce structured data suitable for further processing. Based on extracted categories, links to Wikipedia entries, grouping of similar articles into stories and linking across languages, trends can be detected and spikes reported. A description of the most relevant products built on top of NewsFeed follows.

EventRegistry, a system that groups articles into structured events and allows their exploration. It interlinks articles written in different languages, thus assigning them to the same event. For each event, core information is extracted, such as event location, date, who is involved and what it is about. Users can explore events using extensive search options, visualize and aggregate search results, inspect individual events and identify related events (see [5]).

3XL News is a mobile (iOS) application that provides instantaneous global news monitoring and analysis across several languages on-the-go. It shows how semantic technologies are used in a real-world scenario. 3XL News "Medical" adaptation shown in Figure 6 monitors and provides real-time analysis of health-related news (see [5]).

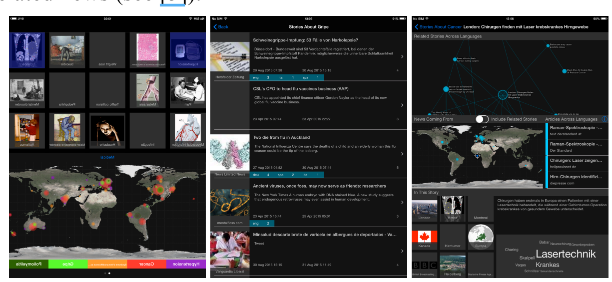

Figure 6 – 3XL news screenhosts providing the stream of news related with a specific topic, their location and the respective relation graph.

3.4 Social Media

Social media mining refers to data mining of content streams produced by people through interaction via the Web. Although social media data are noisy, dynamic and unstructured, mining social media has been found useful for solving a number of research tasks [1,2,7]. For example, Twitter appears to be an efficient tool from the nowcasting point of view [1,8]. The knowledge about secondary attack rates in the influenza season is of importance to assess the severity of the seasonal epidemics of the virus, estimated recently with information extracted from social media in [15]. Figure 7 presents our real-time infographics based of 1% stream of Twitter data for the health domain, displaying keyword, hashtag and user clouds.

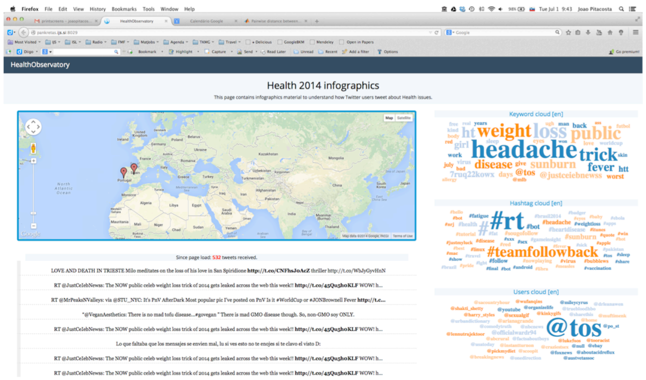

Figure 7 – The infographics based on the 1% stream of Twitter data based on the Twitter observatory, filtered to capture information on the awareness of Influenza and ILI.

4. FUTURE PROJECTS

4.1 Influenza.NET Slovenia

We believe that Deployment of InfluenzaNet to Slovenia would be an important step towards fulfilling the Slovenian E&M Health Strategy as outlined in Section 2. Technical issues, like translation and localisation, comprise a smaller part of the task. Experience from the deployment in other countries has shown that efficient PR activities and the involvement of national public health authorities are essential for the success of the project.

4.2 BigData for Public Health

Information obtained from state-of-the-art media analysis tools developed by AILAB could support InfluenzaNet predictions or be used on InfluenzaNet web sites as additional information for users and, ultimately, prompt health authorities to take corrective actions. We postulate that analysing tweets could further complement results obtained from Influenza.net questionnaires. Analysing social media streams by keywords related to health and epidemic domain allows for correlation with classical epidemic data and for learning social media epidemic patterns. External information, such as weather or major socio-economic events influencing the spread of epidemics, can be aligned with social media information. However, a number of challenges, such as informal character of data, user biases and similar should be accounted for while dealing with disease analysis through social media. State-of-the art knowledge in BigData, (social) media, language technologies and public health issues gives AILAB and its partners a competitive advantage in this area.

4.3 Chronic Diseases and the Environment

Relation between respiratory diseases, both chronic and contagious, and several environmental factors has been studied. There are also studies correlating infectious and chronic respiratory diseases. However, to the best of the authors' knowledge, there is no comprehensive all-encompassing study, which would aim to study correlation among all mentioned factors as well as information obtained from personal sensors, wearable devices and microenvironment monitors. In cases where fragmented studies exist, they rarely go beyond traditional statistical approaches. We would propose a large-scale research project, where a multitude of available measurements are processed, BigData approaches used, state-of-the art analytical methods applied (machine-learning, deep-learning, TDA) and feedback-loops tested in properly set-up clinical trials. In the first phase, patients with COPD, asthma and allergic rhinitis would be studied.

5. CONCLUSION

The potential of the novel methods in machine learning and data analysis can provide a great contribution both to Healthcare and to Public Health. Epidemiology, in particular, can take profit of advanced analytic methods that complement classical tools and can provide a global picture of the evolution of a disease. AILAB with its network of scientific and industrial partners can provide off-the-shelf and on-demand tools to be applied in the broad area of E&M Health.

6. ACKNOWLEDGMENTS

The authors would like to acknowledge the support of the EU projects TOPOSYS (FP7-ICT-318493-STREP), XLike (FP7-ICT-2011-288342), SYMPHONY (Contract no: 611875), and EPIWORK (Contract no: 231807).

7. REFERENCES

[1] J. Bollen, H. Mao and X. Zeng. Twitter mood predicts the stock market. Journal of Computational Science 2.1 (2011): 1-8.

[2] K. Bontcheva et al. TwitIE: An Open-Source Information Extraction Pipeline for Microblog Text. In: RANLP. 2013: 83-90.

[3] G. Carlsson (2009). Topology and data. Bulletin of the American Math. Soc. 46.2 : 255-308.

[4] J. M. Chan, G. Carlsson and R. Rabadan (2013). Topology of viral evolution. Proc. Nat. Acad. of Sc.110.46: 18566-18571.

[5] G. Leban, B. Fortuna, J. Brank, M. Grobelnik, Event registry: Learning about world events from news, in: Proceedings of WWW Companion '14, 2014, pp. 107–110.

[6] J. Linge, J. Mantero, F. Fuart, E. Belyaeva et al. Tracking Media Reports on the Shiga toxin-producing Escherichia coli 0104:H4 outbreak in Germany. – Proc. of the eHealth (2011)

[7] L. F. Lopes et al. Automated social network epidemic data collector. INForum informatics symposium. Lisboa. 2009.

[8] I. Novalija, M. Papler, D. Mladenic. Towards Social Media Mining: Twitterobservatory, SIKDD 2014.

[9] D. Paolotti et al (2014). Web-based participatory surveillance of infectious diseases: the Influenzanet participatory surveillance experience. Clinical Microbiology and Infection 20.1: 17-21.

[10] J. Pita Costa and P. Škraba (2014). A topological data analysis approach to epidemiology. European Conference of Complexity Science 2014.

[11] J. Pita Costa and P. Škraba (2015). Topological epidemiological data analysis. ACML Health 2015.

[12] M. Salathe et al. Digital epidemiology. PLoS Comput Biol 8.7 (2012): e1002616.

[13] M. Trampuš, B. Novak, The internals of an aggregated web news feed, in: Proceedings of IS '12, 2012.

[14] S. P. Van Noort et al. Gripenet: an internet-based system to monitor influenza-like illness uniformly across Europe. Euro surveillance, European communicable disease bulletin 12.7 (2007): E5-6.

[15] E. Yom-Tov et al (2015). Estimating the Secondary Attack Rate and Serial Interval of Influenza-like Illnesses using Social Media. Influenza and other respiratory viruses. DOI:10.1111/irv.12321