Situation Awareness in Crowdsensing for Disease Surveillance in Crisis Situations

Peter Haddawy
Faculty of ICT
Mahidol University
Thailand
+66 2 441 0909
peter.had@mahidol.ac.th

Lutz Frommberger
Capacity Lab
Universität Bremen
Germany
+49 421 218 64199
lutz@informatik.uni-bremen.de

Tomi Kauppinen
School of Science
Aalto University
Finland
+358 9 47001
tomi.kauppinen@aalto.fi

Giorgio De Felice
Capacity Lab
Universität Bremen
Germany
defelice@informatik.uni-bremen.de

Prae Charkratpahu
Sirawaratt Saengpao
Phanumas Kanchanakitsakul
Faculty of ICT
Mahidol University
Thailand
{prae.cha, sirawaratt.sae, phanumas.kan}@student.mahidol.ac.th

ABSTRACT
Crowdsensing can provide real time and detailed information about rapidly evolving crisis situations to facilitate rapid response and effective resource allocation. But while challenges such as heterogeneity of data content and quality, asynchronicity, and volume call for robust data integration and interpretation capabilities, situation awareness in crowdsensing for crisis management remains a largely unexplored area of research. In this paper we extend the mobile4D smartphone-based disaster reporting and alerting system with a situation awareness data interpretation and integration layer and demonstrate its application to the problem of tracking cholera outbreaks. The communication workflow in mobile4D-SA supports interaction between crowdsensed information, system predictions, and multifaceted communication between authorities and affected people on the ground.

Categories and Subject Descriptors
H. [Information Systems]. H.5.3 [Group and Organization Interfaces]. I.2 [Artificial Intelligence]. I.2.1 [Applications and Expert Systems]

General Terms: Algorithms, Human Factors.

Keywords: Crowdsourcing, Situation Awareness, Disease Surveillance, Bayesian Networks, Hidden Markov Models.

1. INTRODUCTION
The ubiquity of mobile and smart phones and the proliferation of social media channels has made crowdsourcing a promising approach to sensing, particularly in crisis situations such as natural disasters and disease outbreaks. Crowdsensing can provide real time and detailed information about rapidly evolving situations and can facilitate effective use of scarce resources by identifying important hot spots to which to send disaster or healthcare workers. This is particularly important when dealing with low resource environments and remote locations. But the large volumes of data, the heterogeneity of its content and quality, as well as the asynchronous nature of the information received pose great challenges in making effective use of such data. Without a significant data interpretation layer, useful information may simply get lost in the noise. Yet situation awareness in crowdsensing for crisis management remains a largely unexplored area of research. Indeed, in a recent review of crowdsensing systems for crisis management, Salfinger et al. [12] concluded that the core situation awareness functions of integrated perception, comprehension, and projection remain unsupported by existing systems.

In this paper we present an approach to using Bayesian networks for situation assessment in crowdsensing for crisis situations. In particular, we extend the mobile4D smartphone-based disaster reporting and alerting system [5] with a situation awareness data interpretation and integration layer and demonstrate its application to the problem of tracking cholera outbreaks. The extended mobile4D-SA system is able to integrate information from authority-sensed data and crowdsensed data, infer information about hypotheses of interest from individual reports and background information, and integrate information across multiple reports over time to assess the current situation as well as to project likely future situations. The communication workflow in mobile4D-SA supports interaction between crowdsensed information, system predictions, and multilateral communication among and between authorities and affected people on the ground.

ICTD '15, May 15 - 18, 2015, Singapore, Singapore

ACM 978-1-4503-3163-0/15/05 $15.00
http://dx.doi.org/10.1145/2737856.2737879

2. RELATED WORK
Two notable crowdsourcing systems in the area of Humanitarian Aid and Disaster Relief that provide some form of situation awareness are HADRian [13] and ESA [14]. In contrast to most systems that focus on map-based visualizations of extracted events or stories, HADRian [13] enriches information from extracted terms by mapping them to an ontology, thus providing

increased flexibility in responding to user queries. ESA [14] provides some degree of projection by classifying tweets according to whether they refer to infrastructural damage or to requests for help, but no more integrated summary of the data.

There has been a surge of interest in the application of crowdsourcing for disease surveillance. Approaches include analysis of search queries [2] and micro-blogging activity [11] as well as platforms for users to actively report disease [4]. Analysis in these systems is carried out by examining temporal and spatial distributions of relevant data points as well as social relations and locations in the case of micro-blogging activity. A major issue is screening of relevant from irrelevant data. Approaches to this consist of screening with statistical techniques and learned classifiers [2, 11] and screening by human analysts [4], with the latter having challenges of scalability.

A number of researchers have made use of Bayesian networks in biosurveillance applications. Cooper et al. [3] describe the use of Bayesian networks to model spatio-temporal patterns for non-contagious diseases that can cause outbreaks in a population such as may occur in bioterrorist attacks. Their PANDA system uses health surveillance information such as emergency department chief complaints for individuals in a population to infer the probability of outbreak. Burkom et al. [1] present an approach to using Bayes nets to fuse environmental and human health information to infer the probability of waterborne disease outbreaks. Their network takes as input daily counts of emergency department patient visits, filtered by chief complaint. Mnatsakanyan et al. [10] present an approach to using Bayesian networks to emulate epidemiologists' cognitive analyses of statistical anomalies to detect patterns suggestive of disease outbreaks. Using the output from algorithms for temporal data anomaly detection, their Bayes nets infer the probability of occurrence of disease outbreaks.

3. MOBILE4D SYSTEM DESCRIPTION AND COMMUNICATION MODEL

Mobile4D [5] is a disaster reporting and alerting system for smartphones in which bi-directional communication is a key feature. Mobile4D connects stakeholders on administrative levels and ground level and encourages continued communication. When officials send out alerts or add information such as updates or advice, local stakeholders are notified via a Push message. Mobile4D users receive alerts based on the locations of their devices. In this way notifications can be targeted so that only people in affected areas and the corresponding administrative structures are notified. Furthermore, contact can be established through the app or through traditional means such as SMS or phone calls. Mobile4D is a crowdsourced system so that users on the ground level are able to report hazards and threats like floods, fires, blocked roads, and diseases via a smartphone app. The information is then sent over the Internet to authorities on the appropriate administrative level. Ground level reports are also distributed to other local stakeholders to encourage communication and ground level solutions. A key strength of mobile4D is a continuous flow of information between affected stakeholders, making mobile4D an ideal platform to provide situation awareness and integrate dissemination of health information into the system workflow. We call the mobile4D system with the situation awareness extension mobile4D-SA.

In January 2015 mobile4D was deployed as a pilot system in three provinces of Lao PDR: Attapeu, Salavan, and Xekong. Provincial and district staff of the Ministry of Agriculture and Forestry have been trained and system use is being monitored in order to optimize workflow and performance under real conditions.

Let us consider an example of the communication flow. With the assessment of risk factors and first reports on disease symptoms, mobile4D-SA can infer the risk of a cholera outbreak. Officials in the affected region are then notified and can use mobile4D-SA to send relevant information to local people such as general advice on hygiene and, most importantly, information of symptoms of cholera with the invitation to report observed symptoms back to mobile4D-SA. If then mobile4D-SA, based on new reports and risk factors, predicts a cholera outbreak, officials may send out an alert with updated information. In parallel, the disease outbreak prediction module of mobile4D-SA will predict the severity and spread of the disease, and in a next step notify further authorities and local people. This communication workflow extends classical situation awareness and crowdsensing systems by mutual interaction between system-made predictions and crowdsensed information and the opportunity for rich communication between authorities and affected people on the ground.

4. SITUATION ASSESSMENT ARCHITECTURE

Mobile4D-SA follows the general architecture of the JDL data fusion model [9, 12]. The Sensing level (L0), consisting of report receipt and feature extraction, receives raw reports from mobile devices and tags them with a time and geo-reference coordinate. Reports are then grouped according to regions. While the system architecture can handle general hierarchical regions such as villages, districts, and provinces, the current implementation divides space into a 10 km by 10 km grid (as also used for distribution of notifications in mobile4D) and groups together reports that fall in a common grid square. In the feature extraction step, keywords and phrases relevant to the domain of interest are extracted from the report.

The perception (L1) and comprehension (L2) levels are handled by a process that interprets each received report. This is done using a diagnostic Bayesian network. The network fuses information from the report with information about general background factors, such as environmental factors. The keywords extracted from each report and the background factors are entered as evidence into the Bayes net and the likelihoods of hypotheses of interest are inferred.

The collection of information inferred from the reports is aggregated and then used by the projection level (L3), which consists of a time-sliced Bayesian network. The network represents time discretely and takes as input information about background factors as well as information inferred from the diagnostic Bayes net at the previous level. It produces projections in terms of likelihoods of future states of hypotheses of interest. Projection results are graphically displayed on a map in relation to the previously defined geographic regions and alerts are sent to relevant channels.

5. CHOLERA DISEASE SURVEILLANCE

This section presents an instance of the mobile4D-SA architecture implemented as a prototype system for cholera disease surveillance. This domain was chosen because cholera is commonly associated with flooding in developing countries.

5.1 Report Interpretation

Reports consist of self-reported illness symptoms that individuals are experiencing. In an examination of a variety tweets related to disease outbreaks, Krieck et al. [7] found that self-reported symptoms are the most reliable information in determining relevance to an outbreak or not since even if people do not know what their problem is, they can readily write about how they feel. Each report is interpreted using the diagnostic network shown in Figure 1. The network contains a central hypothesis node for the severity of cholera, as well as a node indicating simple presence or absence of cholera. The latter has the state Yes if cholera is either mild, moderate, or severe and is used as the main interpretation of the report. Cholera is influenced by environmental risk factors such as contaminated water. Cholera has the direct symptoms vomiting, diarrhea, and rice-water stools. Vomiting and diarrhea can cause dehydration with noticeable symptoms low blood pressure, cold skin, thirst, shock, sunken eyes, weak pulse, weakness, muscle cramps, and reduced urine. Links in this network and in the outbreak projection network are quantified with conditional probabilities. Probabilities were obtained from published statistical studies, e.g [6,8].

For each report received, symptoms are entered using extracted keywords. For symptoms not mentioned, the Bayes net simply reverts to the prior probability. Environmental risk factors are entered based either on report information or from conventional information sources, e.g. government surveillance data. For each report in a given grid square on a given day, the probability of that person having cholera is computed and the number of reported cases is taken to be the expected value of the probabilities computed over all reports received in that day in that grid square.

5.2 Disease Outbreak Projection

Disease outbreak projection is performed by a separate epidemiological Bayes net, shown in Figure 2. The time-sliced Bayes net represents a hidden Markov model of the number of sick individuals in the population over time. There is one copy of the basic network structure for each day with links connecting the time slices. The network is contextually generated dynamically depending on the number of days for which the projection is needed and the population of the region under consideration. The hidden hypothesis nodes are New_Cases and Total_Cases.

Evidence for the number of new cases is provided by reported cases of cholera symptoms. Since not every sick individual will report their symptoms, we assume that the number of reports received is an underestimate, which is quantified in the conditional probability of the number of reports given the actual number of new cases. Since the exact correlation between cases and reports will vary depending on the prevalence of mobile phones and the reporting behavior of the particular population, the value can be set based on experience. The value for the number of reported cases is simply the expected value of the probabilities for the Cholera_Present node in the diagnostic network over all reports received in the current day.

The number of new cases is causally influenced by the number of cases in the previous day and this causal link is quantified in terms of the infection rate of the disease. The New_Cases node also has a link from the Environment node, representing the fact that the infection rate is attenuated by the state of the environment so that, for example, if the sanitation condition is poor the infection rate is higher than average. We assume that once someone contracts cholera, they continue to have it, so that the Total_Cases in a day is simply the sum of the number of cases in the previous day and the number of new cases.

5.3 Linking to External Data with RDF

Storing and structuring information on diseases and their characteristics is essential to building robust models and supporting communication. For this we make use of the Resource Description Framework (RDF) which allows describing not only textual and numerical information about a disease but also linking to useful resources in online Linked Open Data. For instance, linking the description of cholera to an online resource (http://live.dbpedia.org/resource/Cholera) enriches the description by links to related subjects (like pandemics or waterborne diseases) and different language versions. These support enriched functionality such as multilingual reporting and alerting. A Linked Open Data architecture also supports fusion of information from crowdsourced reports with background information, as well as sharing of situation assessment results.

6. EXAMPLE

We demonstrate the situation assessment capabilities with a simulated example of disease due to flooding over the period September 8 - 11. Suppose two reports of disease symptoms are received from a particular cluster of villages on Sept 8, three on Sept 9 and five on Sept 10. We also have information that the water in that area is contaminated throughout the period. The two reports received on Sept 8 are shown in Table 1. Figure 1 shows that entering the information from the first report and the environmental information into the diagnostic network yields a probability of cholera of 0.96. The data from the other report is similarly entered into the diagnostic network, yielding a probability of 0.94. Mobile4D-SA then calculates the expected number of reported cases of cholera as the sum of the probabilities: 1.90. This value is used as the value of the reports_t1 node in the epidemiological network, as shown in Figure 2. Similarly, the expected number of reported cases for days t2 and t3 are entered, as well as the environmental conditions for the three days. After propagating the evidence, we obtain the projected total cases of cholera on Sept 8, 9, 10, as well as Sept 11. The assumption that the reports received may under report the number of actual cases is reflected in the probabilities of new_cases_t1 in the network in Figure 2. Inferred information about the number of projected cases is then displayed on a map of the region (not shown).

Table 1: Reports received on Sept 8 and inferred probabilities

Date	Symptoms	Location	Env. Risk Factors	P(Cholera)
08-09-14	vomiting, diarrhea, muscle cramps	latitude: 46 longitude: 14	contaminated water	0.96
08-09-14	diarrhea, reduced urine, sunken eyes	latitude: 46 longitude: 14	contaminated water	0.94

7. CONCLUSIONS

By combining a crowdsourcing-based communication model with a situation awareness layer, Mobile4D-SA provides a platform for rapid identification and response to emerging crisis situations. The system engages citizens in detection of disease outbreaks by collecting and interpreting their reports of disease symptoms. Reliability of interpretation of reports is increased by fusing report information with information on prevailing environmental risk factors. The system provides a coherent picture of emerging situations to decision makers by integrating report information into an epidemiological model of the disease to produce projections that are graphically displayed on maps.

The current system is yet a prototype and there remain a number of directions for further development and research. The accuracy of the model needs to be validated against data from an actual cholera outbreak. The epidemiological model provides projection within each geographic region but does not yet include

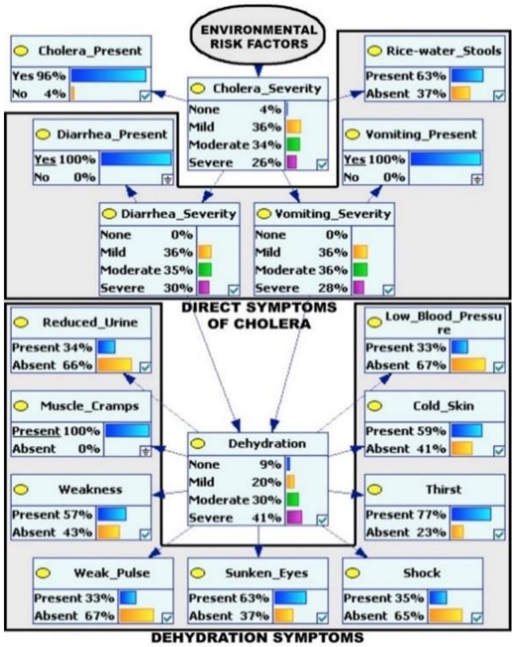

Figure 1: Diagnostic Network with evidence entered from one report.

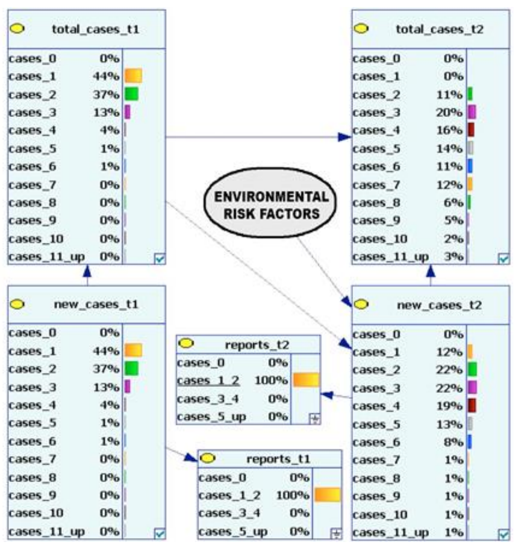

Figure 2: Epidemiological Network. Evidenced entered is: reports_t1 = cases_1-2, reports_t2 = cases_1-2, reports_t3 = cases_3-4, and contaminated_water = Present three time slices (only two shown).

transmission between geographic regions. Finally, there is great promise in using Bayes nets to create generic situation awareness structures that can be assembled and parameterized in response to particular circumstances. Because the structure of the Bayesian network models nicely mirrors that of the physical phenomenon being modeled, this is a promising direction for further development. The use of Linked Data technologies can support such flexible modeling by providing access to information about diseases such as incubation times and information about relevant geographic features such as waterways in a standard machine parsable format.

8. ACKNOWLEDGEMENTS

This work was partially supported by the German Research Foundation through the Collaborate Research Center SFB/TR 8 ”Spatial Cognition”. We thank the Ministry of Agriculture and Forestry of Lao PDR for their longstanding collaboration and the mobile4D student project participants for their contributions.

9. REFERENCES

[1] Burkom, H.S., Ramac-Thomas, L., Babin, S., Holtry, R., Mnatsakanyan, Z., and Yund, C. 2011. An integrated approach for fusion of environmental and human health data for disease surveillance, Statistics in Medicine, 30:470-479.

[2] Chan, E.H., Sahai, V., Conrad, C., Brownstein, J.S. 2011. Using web search query data to monitor dengue epidemics: A new model for neglected tropical disease surveillance, PLoS Neglected Tropical Diseases, 5(5), e1206.

[3] Cooper, G.F., Dash, D.H., Levander, J.D., Wong, W., Hogan, W.R., and Wagner, M.M. 2004. Bayesian Biosurveillance of Disease Outbreaks, In Proc. of the 20th Conf. on Uncertainty in Artificial Intelligence (UAI '04). AUAI Press, Arlington, 94-103.

[4] Freifeld, C.C., Chunara, R., Mekaru, S.R., Chan, E.H., Kass-Hout, T., Iacucci, A.A., and Brownstein, J.S. 2010 Participatory Epidemiology: Use of Mobile Phones for Community-Based Health Reporting. PLoS Medicine, 7(12): e1000376.

[5] Frommberger, L. and Schmid, F. 2013. Mobile4D: Crowdsourced Disaster Alerting and Reporting, In Proceedings of the Sixth Int’l Conf. on Information and Communications Technologies and Development (ICTD '13), Vol. 2. ACM.

[6] Jackson, B.R., et al. 2013. Seroepidemiologic Survey of Epidemic Cholera in Haiti to Assess Spectrum of Illness and Risk Factors for Severe Disease. The American Journal of Tropical Medicine and Hygiene, vol. 89, no. 4, 654-664.

[7] Krieck, M., Dreesman, J., Otrusina, L., and Denecke, K. A new age of public health: Identifying disease outbreaks by analyzing tweets. Proceedings of Health WebScience Workshop, ACM Web Science Conference, 2011.

[8] Kur, L., Mounir, C., Lagu, J., Muita, M., Rumunu, J., Ochieng, B., Weathers, A., Nsubuga, P., Maes, E., and Rolle, I. 2009. Cholera Outbreak - Southern Sudan, 2007. Morbidity and Mortality Weekly, 58(13), 337-341.

[9] Llinas, J., Bowman, C., Rogova, G., Steinberg, A., Waltz, E., and White, F. 2004. Revisiting the JDL data fusion model II. In Proceedings of the Seventh International Conference on Information Fusion (FUSION 2004), 1218-1230.

[10] Mnatsakanyan, Z.R., Burkom, H.S., Coberly, J.S., and Lombardo, J.S. 2009. Bayesian information fusion networks for biosurveillance applications. Journal of the American Medical Informatics Assoc, 16(6): 855-863.

[11] Sadilek, A., Kautz, H. Silenzio, V. 2012. Predicting disease transmission from geo-tagged micro-blog data. Proc. of the Twenty-Sixth AAAI Conf. on Artificial Intelligence (AAAI-2012), pp 136 - 142.

[12] Salfinger, A., Girtelschmid, S., Proll, B., Retschitzegger, W., and Schwinger, W. 2015. Crowd-Sensing Meets Situation Awareness: A Research Roadmap for Crisis Management, In Proceedings of the 48th Annual Hawaii International Conference on System Sciences (HICSS-2015), Hawaii.

[13] Ulicny, B., Moskal, J., and Kokar, M. 2013. Situational Awareness from Social Media. Eighth Conference on Semantic Technologies for Intelligence, Defense, and Security, Fairfax, VA, USA.

[14] Yin, J., Lampert, A., Cameron, M., Robinson, B., and Power, R. 2012. Using Social Media to Enhance Emergency Situation Awareness. IEEE Intelligent Systems, vol. 27, no. 6, 52-59.