EPI-SPIRE: A SYSTEM FOR ENVIRONMENTAL AND PUBLIC HEALTH ACTIVITY MONITORING

Chung-Sheng Li, Charu Aggrarwal, Murray Campbell, Yuan-Chi Chang, Gregory Glass, Vijay Iyengar, Mahesh Joshi, Ching-Yung Lin, Milind Naphade, John R. Smith, Belle Tseng, Min Wang, Kung-Lung Wu, Philip Yu*

IBM Thomas J. Watson Research Center, P O Box 704, Yorktown Heights, NY 10598
*The Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205

ABSTRACT

Health activity monitoring (HAM) has received increasing attention due to the rapid advances of both hardware and software technologies and strong environmental and public health needs. In this paper, we describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel health activity monitoring system that generates alerts from environmental, behavioral, and public health data sources. A model-based approach is used to develop disease and behavior models from multi-modal heterogeneous data sources. Furthermore, a model-based indexing technique has been developed to speed up the data access and retrieval. This system has been successfully applied to various genuine and simulated diseases outbreaks scenarios[1].

1. INTRODUCTION

Recent advances in both hardware and software technologies enable real-time or near real-time monitoring and alert generation for environmental and public health related activities. Environmental related activities include global climate change (such as global warming), deforestation, natural disaster, forest fire, and air pollution. Monitoring of disease outbreaks for public health purposes based on environmental epidemiology has been demonstrated for a number of vector-born diseases such as Hantavirus Pulmonary Syndrome (HPS), malaria, and Denge fever [1-5]. Recently, health activity monitoring (HAM) concept has also been applied to the early detection of subtle human behavior changes due to disease outbreak to provide advanced warnings before significant casualties registered from clinical sources.

The alerts generated from HAM systems are triggered through the fusion of both traditional and non-traditional multi-modal heterogeneous data sources. Traditional data includes data generated from clinical sources such as in-patient and outpatient data. Non-traditional data sources include those data collected from remote sensing (including satellite images), video/audio surveillance, and other data to enable the possibility of extrapolating human behavior.

In this paper, we describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel HAM system capable of generating early warning from monitoring environmental and public health activities. A model-based approach is used to develop the disease and behavior models from multi-modal heterogeneous data sources. Furthermore, a model-based indexing technique has been developed to speed up the data access and retrieval. This system has been successfully applied to vector-born infectious disease such as HPS, pests in the agriculture area such as fire ants, and influenza. For HPS, the advanced warning for high risk regions by using a combination of satellite images and digital elevation map (DEM) can be as much as 9 months [5]. In the case of influenza, preliminary results indicate that early warnings can be generated by Epi-SPIRE using

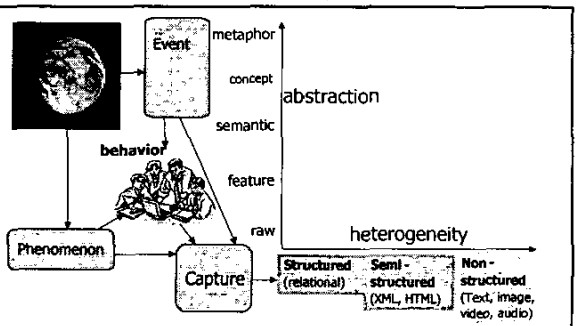

Figure 1: Process of generation of multi-modal heterogeneous data sources.

[1] This research is sponsored in part by the Defense Advanced Research Projects Agency and managed by Air Force Research Laboratory under contract F30602-01-C-0184 and NASA/IBM CAN NCC5-305.
The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied of the Defense Advanced Research Projects Agency, Air Force Research Lab, NASA, or the United States Government.

heterogeneous non-traditional data sources earlier than that can be achieved by using only traditional clinical data sources, thus demonstrating the potential benefit of such a system for public health applications.

2. PRELIMINARY ON HEALTH ACTIVITY MONITORING

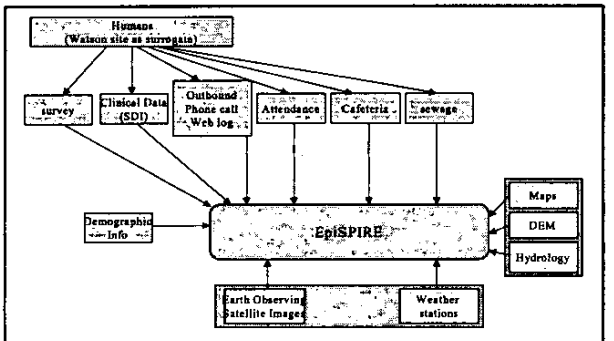

Figure 2: Epi-SPIRE environment.

The multi-modal heterogeneous data sources collected by a HAM system can come from a wide variety of sources, including (1) sensors monitoring the environment either through *in situ* or remote sensing (such as satellites) to capture the events and phenomenon as they occur; (2) data already collected for other purposes, such as e-seminar, phone records, web log, newsgroup, sewage records; (3) data collected from clinical sources such as insurance claims, in-patient and outpatient data, lab tests, and Emergency Room records.

The data sources capturing events and phenomenon related to environments and human behavior, as shown in Fig. 1, can be categorized as structured (parametric or relational), semi-structured (HTML or XML), and non-structured (text, image, audio, and video). The data can be potentially captured at various abstraction levels, including raw data (raw images or video), features extracted from the raw data (such as texture and spectral histogram from

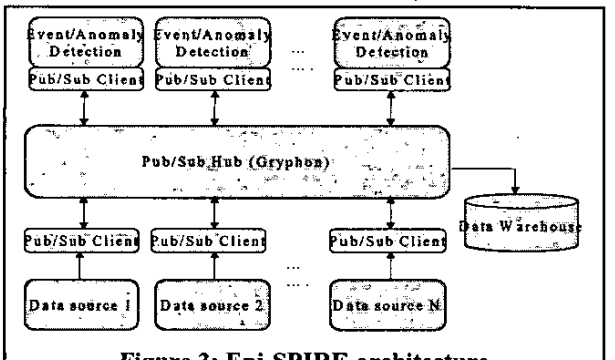

Figure 3: Epi-SPIRE architecture

satellite images), semantic (road, houses), concepts (house surrounded by bushes), and metaphors.

The main challenge in HAM is to be able to fuse multi-modal heterogeneous information sources (based on models) at different abstraction levels, generate multiple hypothesis of the models for the events, phenomenon and behaviors, and test the validity of the hypothesis using the available data. The end objective of such a system is to predict or detect an upcoming event using the model derived from the fused heterogeneous data sources.

3. ENVIRONMENTS AND ARCHITECTURE

The system environment of Epi-SPIRE is shown in Fig. 2. The Epi-SPIRE system uses (1) data collected from the natural environment (such as those collected by the satellites and weather stations), (2) data collected passively as a byproduct of human behavior (such as attendance at work or school, consumption records at cafeteria, sewage generation, web log and phone records), (3) data collected actively from probing the population that are being monitored, usually through periodic survey. In addition to the dynamic data that require real time processing, Epi-SPIRE also utilizes static data such as maps, digital elevation map, hydrology, and demographic information.

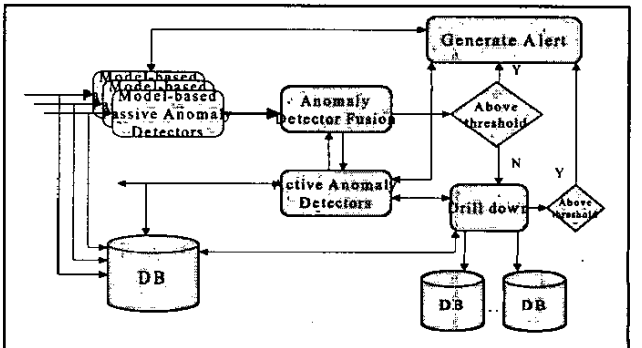

Figure 4: Alert generation process from active and passive detectors.

The system architecture for Epi-SPIRE, which is based on the use of a content-based publisher/subscriber hub - Gryphon [6], is shown in Fig. 3. All of the data sources are connected to the pub/sub hub as publisher so that the data (numeric message, text, audio, or video) from these sources can be routed through the hub to those subscribers that subscribe to these sources. All of the detectors are attached to the system as subscribers as well as publishers, so that they can subscribe to a number of data sources as well as the output from other detectors based on the topics of the data sources.

Note that each of the detectors within the system (as shown in Fig. 3) may generate alerts based on the specific charter of the detector. There is also system level alert generation that fuses the alerts generated from other detectors. The system level alert generation uses alerts

generated by both passive and active detectors, as shown in Fig. 4.

4. MODEL-BASED DATA FUSION AND DETECTION

A number of modeling techniques have been developed in this system to model the spatio-temporal risk factor to certain infectious diseases (HPS, influenza, Denge fever, and anthrax). A linear time-invariant model, $Y = a_1X_1 + a_2X_2 + ... + a_nX_n$, has been used to model the HPS, where each X_i represents the data itself or derived attributes/features from the multi-modal information sources, while the coefficient a_i represents the weights (relative contribution) of the attribute derived from the data. More specifically, the risk assessment model for the risk to HPS associated with a location (x,y) is:

$$R(x,y) = 0.443X_1 + 0.222X_2 + 0.153X_3 + 0.183X_4,$$

where X_1, X_2, and X_3 correspond to the pixel value of band 4, 5 and 7 of Landsat Thematic Mapper image at location (x,y), while X_4 corresponds to the elevation (in meters) from the corresponding DEM (digital elevation map). A risk map based on this model for the south western US during the summer of 1992 is shown in Fig. 5. The actual HPS outbreak took place in 1993 with more than 85% of the cases occur within those highest risk areas. In addition to the linear model, finite state machine models have been successfully developed and applied to modeling the risk to fire ants (which are harmful to both crops and livestock of the southeast US), and Bayesian network models have been developed for other infectious diseases.

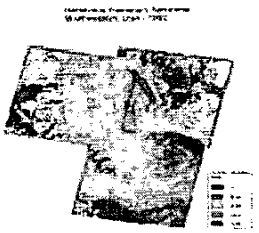

Figure 5: Risk map for Hantavirus Pulmonary

The same model for data fusion can also be used for indexing to facilitate model-based information retrieval. A model-based indexing technique, Onion [7], was developed for linear model based data fusion and retrieval and provide up to three order-of-magnitude speedups as compared to linear evaluation.

The risk map generated above provides the baseline for anomaly detection – as we are usually only interested in unexplainable anomalies. We have explored two general classes of model-based anomaly detectors (Fig. 3 and 4) that have applicability to site surveillance. The first class, which we term differential detectors, is applicable in the case where there are two or more sites that have similar behaviors. A differential detector raises an alarm when the deviation between sites becomes sufficiently large. The second class of detectors is predictive, i.e., they predict "normal" site behavior and raise an alarm if a sufficiently large deviation from normal is detected.

5. VALIDATION

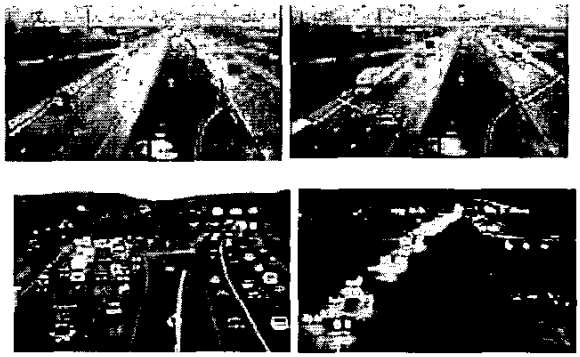

Figure 6: car counting for monitoring traffic.

The Epi-SPIRE system has been validated in a genuine environment between the fall 2001 and summer of 2002 to monitor the behavioral changes of a population caused by the earliest stages of illness. Examples of such behaviors include increased absenteeism, increased inquiries for medical information, changes in eating/drinking habits, increased coughing, increased traffic for leaving the building early, and increased sewage generation. IBM T. J. Watson Research Center, which consists two sites - Yorktown and Hawthorne, and is located in Westchester County, NY (50 km north of New York City), is used in this case study. The total population for the sites is approximately 2000. All of the data collected below have been properly anonymized so that the privacy of the population being investigated is not violated.

1) A weekly survey of self-reported health level was conducted from January 2002 through May 2002, during which an email-based survey of the population was run at the Watson site. About 400 IBM employees volunteered to participate. This survey had an excellent response rate: 92% of polled employees responded the same day, 73% by noon.
2) The IBM Watson worksite requires the swiping of a badge in order to gain entry. The badge number and time of entry are recorded in a database that is maintained for security purposes. We have been receiving an anonymized version of this information since 12/2001.
3) The IBM Watson site records, for billing purposes, all phone calls made outside the site. The calling number, called number, time of call, and duration of

call are recorded in a database. A set of local medically related phone numbers was obtained from two main sources (scanned from yellow pages, internet directories). From an anonymized version of these data it is possible to count the number of calls made from Watson to medically related numbers, as well as the number of extensions that were used to place these calls.

4) The IBM Watson site records, for security purposes, all accesses to external websites at the firewall. The source IP, destination IP, and date/time of access are recorded in a database. Using an anonymized version of this database along with a manually generated list of medically related websites, it is possible to count the number of accesses to these medically related sites, as well as the number of computers from which these requests were made.
5) Consumption of cafeteria food and beverages at Hawthorne Cafeteria (one of the two sites for the IBM T. J. Watson Research Center) are recorded electronically. This cafeteria provides service to about 700 people.

	Total Frames	Inbound (A/M)	Outbound (A/M)	Precision
Clip 004	2417	106/95	106/100	**91.3%**
Clip 007	3856	33/32	45/44	**97.4%**
Clip 021	5460	143/143	62/65	**98.6%**
Total	11733	282/270	213/209	**96.7%**

Figure 7: Precision of car counting.

6) A number of other potential data sources have been considered and undergone some preliminary evaluation. These include: site utility usage, site sewage generation, cough counting, and car counting (cars entering or leaving site). Specifically, the car counting is based on the use of the video captured from the webcam (shown in Fig. 6) in order to capture potential early departure traffic from a site. The car counter is fairly accurate except during the night or when it is raining, as shown in Fig. 7 [8].

The alerts generated from these data sources are compared to the insurance claims from the Westchester County. There is preliminary evidence that the warnings generated by some of the data sources (survey and phone in particular) lead the clinical sources.

We have also evaluated the Epi-SPIRE anomaly detection mechanisms in a synthetic environment in which site-specific or regional outbreaks are simulated. The results indicate that the pathogen release can be detected within 4 days for acceptable false alarm levels.

6. SUMMARY

In this paper, we describe the architecture and implementation of the Epi-SPIRE prototype, which is a novel health activity monitoring (HAM) system that generates alerts from environmental, behavioral, and public health data sources. A model-based approach is used to develop the disease and behavior models from multi-modal heterogeneous data sources. This system has been successfully validated in a number of scenarios involving infectious disease outbreak.

7. REFERENCES

[1] Glass, GE, T. L. Yates, J. B. Fine, T. M. Shields, J. B. Kendall, A. G. Hope, C. A. Parmenter, C.J. Peters, T. G. Ksiazek, C.-S. Li, J. A. Patz and J. N. Mills. "Satellite imagery characterizes local animal reservoir populations of Sin Nombre virus in the southwestern United States," Proc. National Academy of Science 99:16817-16822. (December 23, 2002)

[2] Glass, G. E. "Public health applications of near real time weather data". 6th Earth Sciences Information Partnership Conf. 2001.

[3] Klein, S. L., A. L. Marson, A. L. Scott, GE Glass, "Sex differences in hantavirus infection are altered by neonatal hormone manipulation in Norway rats," Soc Neuroscience 2001.

[4] Klein, S. L. , A. L. Scott, G. E. Glass, "Sex differences in hantavirus infection: interactions among hormones, genes, and immunity," Am Physiol Soc. 2001.

[5] Glass G. E. "Hantaviruses. Climate Impacts and Integrated Assessment," Energy Modeling Forum 2001.

[6] S. Bhola, R. Strom, S. Bagchi, and Y. Zhao, "Exactly-once Delivery in a Content-based Publish-Subscribe System," Dependeable Systems and Networks 2002.

[7] Y.-C. Chang, L. D. Bergman, V. Castelli, C.-S. Li, M.-L. Lo, and J. R. Smith, "The Onion Technique: Indexing for Linear Optimization Queries," ACM SIGMOD 2000, May, 2000.

[8] Belle L. Tseng, Ching-Yung Lin, and John R. Smith. Real-Time Video Surveillance for Traffic Monitoring Using Virtual Line Analysis IEEE ICME, Lausanne, Switzerland, August 2002.