Monitoring Food Safety by Detecting Patterns in Consumer Complaints

Artur Dubrawski[1], Kimberly Elenberg[2], Andrew Moore[1] and Maheshkumar Sabhnani[1]

(1) The Robotics Institute, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213
Artur.Dubrawski@cs.cmu.edu, Andrew.W.Moore@cs.cmu.edu, Maheshkumar.Sabhnani@cs.cmu.edu
(2) Food Safety Inspection Service, USDA, 1400 Independence Ave. S.W., Washington, D.C. 20250, Kimberly.Elenberg@fsis.usda.gov

Abstract

EPFC (Emerging Patterns in Food Complaints) is the analytical component of the Consumer Complaint Monitoring System, designed to help the food safety officials to efficiently and effectively monitor incoming reports of adverse effects of food on its consumers. These reports, collected in a passive surveillance mode, contain multi-dimensional, heterogeneous and sparse snippets of specific information about the consumers' demographics, the kinds, brands and sources of the food involved, symptoms of possible sickness, characteristics of foreign objects which could have been found in food, involved locations and times of occurrences, etc. Statistical data mining component of the system empowers its users, allowing for increased accuracy, specificity and timeliness of detection of naturally occurring problems as well as of potential acts of agro-terrorism. The system's main purpose is to enhance discovery and mitigation of food borne threats to public health in the USDA Food Safety Inspection Service regulated products. As such, it is being envisioned as one of the key components of the nationwide bio-security protection infrastructure. It has been accepted for use and it is currently going through the final stages of deployment. This paper explains the motivation, key design concepts and reports the system's utility and performance observed so far.

Introduction

Challenges of Food Safety. The United States agriculture generates more than one trillion dollars worth of economic activity and over 60 billion dollars in food exports (GAO 2005). Unfortunately, the agriculture and food systems are vulnerable to disease, pest, or poisonous agents that occur naturally, are unintentionally introduced, or are intentionally delivered by acts of terrorism. In the US, there are 76 million recorded cases of food borne illness occurring every year, and about 5,000 of them are terminal (Mead et al. 1999). Costs of food borne illnesses or injuries adversely impact the food industry, households, and health sector. The attacks on September 11th, 2001, brought special attention to vulnerability of the US food supply system. It is an extensive, open, interconnected, diverse, and complex structure providing potential targets for terrorist attacks which could have catastrophic health and economic effects.

Consumer Complaint Monitoring System. In order to provide the best possible protection against all those threats, the USDA Food Safety Inspection Service (FSIS) is responsible for numerous initiatives aimed at improving food supply safety and security, such as monitoring and surveillance activities, which require real time tracking and assessment of data in an effort to serve as early warning systems. A significant example is the implementation of an adverse event consumer based surveillance system for food safety and food security referred to as the Consumer Complaint Monitoring System (CCMS). Located within the FSIS Office of Public Health and Safety, the CCMS has evaluated nearly 4,000 consumer complaints from January 2001 through January 2005. Results of these evaluations have led to recalls of unwholesome adulterated products, improved quality assurance, and analyzes of hazards and critical control points in slaughter and processing plants.

The CCMS is an intranet electronic database used to record, evaluate, and track all adverse food events involving specifically meat, poultry, and egg products reported to FSIS. The two main portals of entry for consumer complaints are through phone calls to the field Compliance Officers stationed throughout the United States and through the 1-800 phone number of the Meat and Poultry Hotline. Additional data sources include consumer complaints reported by a state or local health department or another federal agency, such as the Food and Drug Administration, as well as complaints that involve imported products that have been re-inspected at the port of entry. Most of the consumer complaints involve: illnesses that occurred after eating a product; injuries that occurred while eating a product; foreign objects that were found in a product; allergic reactions that occurred after eating a product; suspected under-processing of ready-to-eat product; allegations of improper labeling of a product; and dissatisfaction with the quality of a product.

Upon arrival of the new complaint case, the FSIS analysts search the CCMS database for similar complaints. Multiple cases bearing similar features reported independently on each other may indicate a possibility of a serious food

safety problem. Analysts initiate an investigation and mitigating actions for all possible events detected that indicate potential health safety, or security issues. Investigation examples include: a laboratory confirmed food borne illness, an alleged allergic reaction due to a previously diagnosed food allergy to an unlabeled ingredient, or signs that a ready-to-eat product may be under-processed. If the CCMS database contains more than one apparently related complaint about a foreign material in a product produced at a particular establishment an investigation is also initiated.

Monitoring Emerging Patterns in Food Complaints with EPFC. The EPFC is the analytic core of the CCMS. It is the first practically applied instance of *Tip Monitor*, a broader concept of a statistical data mining approach to practical alerting from multivariate rare-event data (Sabhnani et al. 2005). The purpose of Tip Monitor is to identify small groups of significantly related records in an incoming stream of event-based data such as anonymous individual patient health events involving chief complaint strings, prescription orders, public safety hotlines or customer complaints. Very often, such data is sparse, noisy and it may contain little and spotty evidence of potentially crucial coincidences. That last feature makes it very hard to detect important events with more traditional approaches used by bio-surveillance analysts such as scan statistics or multivariate time series analysis, since they are designed to benefit from ample evidence (Buckeridge et al. 2003, Neill and Moore 2004, Wong et al. 2003).

Suppose that among the chief complaint strings of two unrelated patients in the same city on the same date there was mention of bloody stools in pediatric cases. The multiple mentions of "bloody stools" or "pediatric" might not be surprising, but the tying together of these two factors, given matching geographic locations and timings of reporting, is sufficiently rare that seeing only two such cases is of interest. This was precisely the evidence that was the first noticeable signal of the tragic Walkerton, Canada, waterborne bacterial gastroenteritis outbreak caused by contamination of tap water in May 2000 (Mackay 2002). That weak signal was spotted by an astute physician, not by a surveillance system. Reliable automated detection of such signals in multivariate data requires new analytic approaches.

Similar circumstances accompany the task of detecting systematic associations between individual food consumer complaints received by the USDA. FSIS analysts need to be alerted even if merely two independently collected complaints seem to be substantially related to each other. Potential relevance must be evaluated using spotty multivariate data subjectively reported by the complainants. Even seasoned analysts agree that it is a tedious task, prone to subjective judgment and human error. This difficulty is augmented by the underlying complexity of data interpretation. For instance, a food borne illness often presents itself with flu-like symptoms such as nausea, vomiting, diarrhea, or fever. That makes it difficult for analysts to determine if illness was caused by bacteria or other pathogens in food and if the pathogens occurred naturally or intentionally, however illness could also be caused by chemicals or heavy metals. The underlying complexity of the task at hand results in highly subjective judgments and it may lead to erroneous decisions.

There is a need for a statistical data mining system capable of automated and objective monitoring of the stream of incoming complaints for evidence of linkages with records from the recent past. It should help the analysts by alerting them early about emerging patterns in food complaints, and by focusing their attention to the most probable linkages. Effectively, it should improve their situation awareness, reliability of decisions and response times. The EPFC is aimed at fulfilling that need.

Approach

The EPFC is designed to screen sparse and noisy data for potential linkages between individual reports of adverse effects of food on its consumers. These reports contain multi-dimensional and heterogeneous snippets of specific information about the consumers' demographics, the kinds, brands and sources of the food they ate, symptoms of sickness they may be experiencing, characteristics of foreign objects which could have been found in food, involved locations and times of occurrences, and so on.

In EPFC, the notion of two cases being similar is determined by a probabilistic model that is partially learned from historical data and partly obtained from experts. This is in contrast to simple pair-wise distance measures and string similarity scoring, which may indicate a correlation but not necessary a factual connection between the two events.

Knowledge of the domain expert is modeled into the form of a list of causal scenarios. Each scenario, such as for instance a malicious contamination of raw food at processing plant, or a product-nonspecific illness occurring in some local community, will focus on a specific subset of features of multivariate records of events, which would be considered relevant to the selected cause. The experts assist in defining the structure and parameters of the probabilistic model of conditional dependencies between features of the pairs of events. In the current version of the system, these dependencies are modeled by a Bayesian network, given a predefined assumed cause. A separate model is constructed this way for each individual scenario of interest. These models, pre-constructed by hand, are combined into one system using Bayes' rule.

Mathematically, the EPFC estimates how likely it is for a newly reported complaint case X_n to be a close copy of some other case in the past data, X_i, if both have been generated by the same specific underlying cause labeled

well as means to collect new entries at comparable frequencies. Integration of those external sources of data should lead to improved reliability of predictions, though it may require additional work into scalability of the EPFC.

Conclusion

The EPFC at its current stage of development already lives up to its promise, as evidenced by testing on a collection of historical food complaints. It is receiving a very positive feedback from its evaluators and future users due to its high utility, sensitivity, and satisfactory accuracy.

It is able to overcome the limitations of typical health safety related event-based data by employing Bayesian techniques to model potential common causes obtained from domain experts' knowledge of scenarios of events triggered by specific causes of interest. In EPFC useful alerts can be generated using more specific information and on fewer cases than typically attainable in syndromic surveillance. In addition, it is sensitive to new emerging patterns of previously unknown or unanticipated adverse events.

There is a potential for secondary benefits from EPFC, going beyond its primary purpose of supporting food safety. They include giving food manufacturers a timely feedback on the safety of their products, which could positively impact stability and sustainable development of local economies which often heavily rely on food industry.

The EPFC is an instance of the more general Tip Monitor concept. As such, it illustrates the ability of this approach to become useful in other domains, where multivariate heterogeneous data comes in a relatively short supply and where early detection of relatively low amplitude signals is required. The natural areas of potential future applications of Tip Monitor include analyzing hospital records for signals of disease outbreaks, analyzing maintenance records for early evidence of systematic patterns of equipment failures, analyzing law enforcement reports, and so on.

Acknowledgement

This work has been performed under the US Government contract (Award #GS06K97BND0710 USDA/FSIS Consumer Complaint Monitoring System II).

References

Buckeridge, D.L.; Burkom, H.; Moore, A.; Pavlin, J.; Cutchis, P.; and Hogan, W. 2003. Evaluation of syndromic surveillance systems – design of an epidemic simulation model, *Morbidity and Mortality Weekly Report*, 2003. 53 (Supplement): 137-143.

GAO 2005. Protecting Against Agroterrorism. Report GAO-05-214, Government Accountability Office, March 2005.

Jacobs, B.; Greenwald, J..; Jain, A.; Park, A.; Wong D.; and Yang, M. 2005. Proposed Cost-Benefit Framework for Bio-surveillance Systems, Auton Lab Technical Report, Carnegie Mellon University, December 2005.

Mackay, B. 2002. Walkerton, 2 years later: Memory fades very quickly", *Canadian Medical Association Journal*, May 14, 2002. 166 (10).

Mead, P.S.; Slutsker, L.; Dietz, V.; McCaig, L.F.; Bresee, J.S.; Shapiro, C.; Griffin, P.M.; and Tauxe, R.V. 1999. Food-Related Illness and Death in the US, *Emerging Infectious Diseases*, 5(5), 1999. (Available from http://www.cdc.gov/ncidod/eid/vol5no5/mead.htm).

Neill D. and Moore A. 2004. Rapid detection of significant spatial clusters, *In Proceedings of the 10th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 256-265.

Sabhnani, M.; Neill, D.; Moore, A.; Dubrawski A.; and W.Wong, W.-K. 2005. Efficient Analytics for Effective Monitoring of Biomedical Security, *In Proceedings of the International Conference on Information and Automation*, Colombo, Sri Lanka, December 2005.

Wong, W.-K.; Moore, A.; Cooper, G.; and Wagner, M. 2003. What's Strange About Recent Events, *Journal of Urban Health*, 80: 66-75.