Monitoring Pharmacy Retail Data for Anomalous Space-Time Clusters

Maheshkumar R. Sabhnani[1], Daniel B. Neill[1], Andrew W. Moore[1],
Fu-Chiang Tsui[2], Michael M. Wagner[2], Jeremy U. Espino[2]
[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213
[2]RODS Laboratory, University of Pittsburgh, Pittsburgh, PA 15213

OBJECTIVE

This paper describes the evolution of a bio-surveillance system that incorporates user feedback to improve system utility and usability. The system monitors national-level over-the-counter (OTC) pharmacy sales on a daily basis. We use fast spatio-temporal scan statistics to detect disease outbreaks.

BACKGROUND

Bio-surveillance systems monitor multiple data streams (OTC sales, Emergency Department visits, etc.) to detect both natural disease outbreaks (e.g. influenza) and bio-terrorist attacks (e.g. anthrax release). Many detection algorithms show impressive results under simulated environments, but the complex behavior of real-world data and high costs associated with processing false positives make it difficult to develop practical bio-surveillance systems. We believe that using expert knowledge from public health officials will help us to better understand the real-world data, improving our ability to distinguish actual disease outbreaks from non-outbreak patterns.

METHODS

The National Retail Data Monitor (NRDM) receives daily sales of ~9000 OTC products (grouped into 18 categories) from over 10,000 stores throughout the nation [1]. Our system searches for spatio-temporal patterns in this data. Given a search region (a city, county, state or even the entire country), the algorithm first maps this search region to a uniform, rectangular $N \times N$ grid. It then searches over all axis-aligned rectangular regions on the grid, in order to find regions that have shown a recent anomalous increase in sales. A detailed description of the algorithm is available in [2-4]. Recent sales in a region are compared with the estimated baselines (calculated using the past 3 months of time series data). Baselines can be estimated using a variety of time series analysis techniques, at various levels of aggregation (store level, grid cell level, or region level) [4]. During baseline estimation, we address missing data issues, weekly trends (e.g. sales on Mondays typically greater than Saturdays), and seasonal trends (e.g. rise in Cough/Cold sales during flu season). The region scoring function assumes that counts follow a Poisson distribution, as in [4], and p-values are calculated using randomization testing. The k-best significant regions are reported as alerts (possible disease outbreaks) to the public health officials.

RESULTS

Our spatial scan statistics (SSS) system reports anomalous regions to users and then incorporates user feedback in order to improve the system's ability to detect outbreaks. Initial versions of the system involved sending alert reports to users via e-mail. To improve users' ability to evaluate outbreaks, alert location maps and a store level data view were added. The salient features in the current system include showing alert-region time series, showing store-level data in the region, and navigating in and around alert regions on the GIS map to help further investigate alerts. Instead of e-mail attachments, users can access alerts online on the SSS website, add and view comments on each alert, and select which alerts they want to see through user defined filters and searches. User feedback has helped us identify false positives due to "single store" increases (bulk purchases or promotional sales) and due to underestimation of baselines, and we have added filters to remove these unwanted alerts, drastically reducing the false positive rate.

CONCLUSIONS

Users have detected a number of unique patterns in OTC data using this system: for example, increases in OTC sales before inclement weather, immediately after a national holiday, and at tourist destinations during long weekends. These patterns underscore the difficulty of determining which increases in sales are due to real outbreaks, and which increases are due to a variety of other unmodeled factors. Continued feedback from users of this system will help us identify true disease outbreak patterns and improve the sensitivity/specificity of our bio-surveillance system.

REFERENCES

[1] Wagner MM, Tsui F-C, et al., A national retail data monitor for public health surveillance. Morbidity and Mortality Weekly Report, 2004, 53 (Supplement): 40-42.

[2] Neill DB, Moore AW, Rapid detection of significant spatial clusters. Proc. 10th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2004, 256-265.

[3] Neill DB, Moore AW, Pereira F, Mitchell T, Detecting significant multidimensional spatial clusters. Advances in Neural Information Processing Systems 17, 2005, 969-976.

[4] Neill DB, Moore AW, Sabhnani MR, Daniel K, Detection of emerging space-time clusters. Proc. 11th ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, 2005, 218-227.

Further Information:
Maheshkumar R. Sabhnani, sabhnani+@cs.cmu.edu
www.autonlab.org