Detecting Anomalous Patterns in Pharmacy Retail Data

Maheshkumar R. Sabhnani
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
sabhnani+@cs.cmu.edu

Daniel B. Neill
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
neill@cs.cmu.edu

Andrew W. Moore
School of Computer Science
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213
awm@cs.cmu.edu

ABSTRACT

This paper describes a bio-surveillance system designed to detect anomalous patterns in pharmacy retail data. The system monitors national-level over-the-counter (OTC) pharmacy sales on a daily basis. Fast space-time scan statistics are used to detect disease outbreaks, and user feedback is incorporated to improve system utility and usability.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Apps—Data Mining

General Terms

Algorithms

Keywords

Cluster detection, space-time scan statistics, biosurveillance.

1. INTRODUCTION

Bio-surveillance systems have recently gained a lot of attention and are growing more and more complex. Multiple sources of data (pharmacy sales, emergency department visits, weather indicators, census information, etc.) are now available, and these sources can be used to identify both natural disease outbreaks (e.g. influenza) and outbreaks resulting from bio-terrorist attacks (e.g. anthrax release). The bio-surveillance research community is actively developing intelligent algorithms to detect outbreaks in a timely manner, in order to save lives and costs. However, though many of these algorithms show impressive results under simulated environments, their performance tends to degrade when applied to real-world datasets. Seasonal and day-of-week trends, missing data, lack of known disease outbreaks, difficulties in designing test beds, and high costs associated with processing false positives are some of the many reasons that hinder development of a successful practical bio-surveillance system. We believe that incorporating expert knowledge from public health officials will provide valuable insight to this complex process of disease outbreak detection. An immediate goal is to provide a tool that not only shows the alarms to the expert users, but also allows them to provide feedback on the alarms. This feedback loop is essential for iterative refinement of outbreak detection tools. This paper highlights our experiences with developing such a bio-surveillance system that currently monitors national level pharmacy sales of over-the-counter (OTC) drugs on a daily basis.

Our system searches for spatio-temporal patterns in the OTC data from pharmacies, grocers and other stores that sell OTC products throughout the United States. Given some search region (which can be 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. The regions that show high deviation in sales from the estimated baselines are labeled as alerts–clusters of OTC sales that may indicate disease outbreaks. A detailed description of the algorithm is available in [1-3]. Given our limited ability to distinguish clusters caused by outbreaks from clusters with other causes, we present selected alerts to public health officials only after they have been filtered by some simple rules to remove unimpressive anomalies. Their feedback is then used to improve the performance of the algorithm. The following sections describe this system in detail.

2. SYSTEM OVERVIEW

The National Retail Data Monitor, developed and operated by the RODS (Real-time Outbreak and Disease Surveillance) Laboratory at the University of Pittsburgh, receives the OTC data from the national and local vendors [4, 5]. The data consists of daily store level sales of 9000 OTC products used for the symptomatic treatment of infectious diseases. The NRDM groups individual product sales into 18 groups of similar products (e.g., Baby/Child electrolytes, Cough/Cold, Thermometers, Stomach remedies, and Internal analgesics). We process the past three months of data (around 5.5 million records) to estimate recent baselines (i.e. the number of sales we would expect to see in each store). Each record includes the store ID, its corresponding zip code, date of sale, and units sold for a particular syndrome. There are more than 10,000 unique stores present in the data. This data is received on a daily basis, with one-day delay from the date of sale. There are various challenges with estimating the store baseline sales. First, there are strong seasonal and weekly trends in the OTC data. Figure 1 shows a sample weekly trend in Baby/Child electrolyte sales. Sales on a typical Monday and Tuesday tend to be higher than on Friday and Saturday. This trend depends on many factors: region location, urban or rural community, etc. Figure 2 shows the seasonal trends in Cough/Cold sales. Average daily sales in the month of March were ~5000 units higher than in April. We have also noticed a sudden rise in sales for days following a national holiday. We address the seasonal and day-

AD-KDD'05, August 21, 2005, Chicago, Illinois, USA.

of-week trends by incorporating them into the baseline time-series analysis. The current data storage schema does not differentiate between missing data (i.e. stores that have not reported sales for a specific date by the time of analysis) and zero counts (i.e. stores that sold zero units on that date). To deal with this limitation, we assume that data are missing only if a store reports no sales for all product categories; if a store has zero counts for some product categories and non-zero counts for others, the zero counts are assumed to result from zero sales rather than from missing data. We infer all missing data points from the time series of counts for that location, using an exponentially weighted moving average technique. Once the time series has no missing data, any reasonable univariate time series algorithm that accounts for day-of-week and seasonal trends can be applied to estimate recent baseline sales.

After we receive the past three months of national OTC data, we define multiple search regions with differing resolution (some states, some counties, and others that cover the entire country). This ensures that we detect large-scale anomalies, and not just daily fluctuations at the store or zip code level. As noted above, the search region is mapped to a rectangular two-dimensional grid of size $N \times N$. We need to know the store locations in order to map them onto the grid cells; however, due to data privacy concerns, we do not have access to the exact longitude and latitude of each store. Instead, we are given the zip code containing each store, and use the longitude and latitude of the zip code centroid to populate the grid cells. The search algorithm then scores every possible axis-aligned rectangular region using the recent baselines (expected counts) and observed counts in the region. Baseline values can be aggregated either for individual stores (the "building-aggregated time series" method, or BATS) for individual grid cells (the "cell-aggregated time series" method, or CATS), or on-the-fly for an entire search region (the "region-aggregated time series" method, or RATS). Additionally, a variety of methods are used for time-series analysis. For details on aggregation techniques and time series algorithms tested on the OTC data, please refer to [3]. The scoring function assumes that baseline sales follow a Poisson distribution. We also perform significance testing on the score of each region by randomization. This helps us remove anomalous regions that could be explained as being generated by chance. The k-best regions (i.e. those significant regions with the highest scores, and therefore the lowest p-values) are reported as possible disease outbreaks.

3. SYSTEM EVOLUTION

The primitive versions (version 1.X) of the current spatial scan statistics (SSS) system involved reporting significant regions via e-mail. Each day, a set of states and counties was scanned for anomalous regions, and the alert results for each state/county were sent as an e-mail attachment to the appropriate public health officials. Though the users were given the latitude, longitude, syndrome, score, and p-value of each alert region, it was difficult for them to get a feel of where exactly the outbreak occurred, or to interpret the probable cause of the alert (i.e. whether it was a real outbreak or a false positive). To deal with these issues, we developed a SSS viewer application tool with a dual purpose. First, it allows end users to browse the data that led to an alert. Second, it provides easy feedback opportunities in which they can tell us which alerts were genuine and which were uninteresting or due to non-outbreak reasons. Figures 4 and 5 show sample screen shots of our viewer tool. Salient features of this tool include showing alert-region time series, showing store-level data in the region, and navigating in and around the alert region on the GIS map to help further investigate the alert. We released this tool during our version 2 release. In this version, all alerts were displayed on the website rather than via e-mails. The current version 3.0 (to be released in June 2005) has enhanced capabilities on the web. Now users can not only view alerts, but they can also rank them, add feedback comments, and give suggestions. Users can also search for alerts using different criteria, such as zip code, score, observed counts, expected counts, etc. We are trying to extract user expertise in identifying features of the clusters that may discriminate between clusters likely due to disease outbreaks and clusters likely due to other causes. Another powerful tool that we have given to users is to add their custom-defined input scripts to the pool of scripts that run daily. Users can set their own grid resolution, change baseline evaluation time series method, set aggregation level, etc. By enabling users to create their own input scripts, we can learn what results and settings are most relevant to real users in the surveillance task. This feedback will help us better manage these alerts and distinguish true outbreaks more efficiently. Figure 3 shows a sample screen shot of the user home page. In the future, we also plan to provide more features (e.g. providing store locations, tracking of previously reported alerts for post analysis purposes, etc.) to the end users so that they can give better feedback.

We have been running this system daily on OTC data for over one year. Initially the algorithm reported a large number of false positives: regions that were statistically significant according to our model but clearly did not correspond to actual outbreaks. Some of these false positives resulted from "single store" anomalies: individual stores with large spikes in sales on a given day. Two possible explanations for these single store anomalies are bulk purchases by a single buyer (e.g. restocking by a hotel, clinic, etc.) or promotional sales. We address this issue by only reporting those regions that have shown increased counts due to multiple stores: in other words, we filter out a region if removing any single store from that region would cause its score to become insignificant. In order to make a simple adjustment for potentially unmodeled fluctuations in day-to-day counts, we also apply a conservative "threshold" filter, which assumes that the baselines were underestimated by some amount (e.g. 15%). If both the "single-store" adjusted score and the "threshold" adjusted score are still significant, we report the region as a potential outbreak. Figure 4 shows a recent potential Baby/Child Electrolyte disease outbreak at the border of Alabama and Georgia. There are 16 stores in this area, and at least five of these stores have shown high deviations from baseline in electrolyte sales. The alert region is not shown in the figure due to data privacy concerns.

We have already observed a number of unique and interesting trends in the OTC data using this system. For example, people tend to buy some products just before inclement weather (such as snowstorms or hurricanes), presumably to stockpile them. There is also typically a rise in OTC sales immediately after a national holiday. Another interesting effect recently observed was increased sales in tourist destinations during long weekends. Figure 5 illustrates this trend during the recent Memorial Day weekend. Since the NRDM has highest coverage in the eastern United States, a large number of tourist destinations (gray highlighted regions on the map) produced alerts resulting from the change in population distribution around these areas. Again, due to data privacy concerns, we have not shown the location of the region whose time series is shown below the map. Although these are interesting results, they 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. In the near future, we intend to increase the number of outbreak indicators: adding more algorithms and data sources to the system. We are planning to add emergency department data and include more independent univariate time-series algorithms to improve our confidence when alerting outbreaks. This system is helping us to understand the real-world OTC data and to improve our detection models and methods. Continued feedback from public health users will increase our ability to differentiate true outbreaks from yet unknown natural causes for increased OTC sales, thus enabling us not only to find "significant" regions, but also to determine which of these clusters are most relevant for public health investigation.

4. ADDITIONAL AUTHORS

Fu-Chiang Tsui, Michael M. Wagner, and Jeremy U. Espino (RODS Laboratory, University of Pittsburgh, Pittsburgh, PA 15213). E-mail: {tsui, mmw, jue}@cbmi.pitt.edu.

5. REFERENCES

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

[2] D.B. Neill, A.W. Moore, F. Pereira, and T. Mitchell. Detecting significant multidimensional spatial clusters. *Advances in Neural Information Processing Systems* **17**, 969-976, 2005.

[3] D.B. Neill, A.W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. Accepted to *11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2005.

[4] M.M. Wagner, F.-C. Tsui, J. Espino, W. Hogan, J. Hutman, J. Hersh, D.B. Neill, A.W. Moore, G. Parks, C. Lewis, and R. Aller. A national retail data monitor for public health surveillance. *Morbidity and Mortality Weekly Report, Supplement on Syndromic Surveillance* **53**, 40-42, 2004.

[5] M.M. Wagner, J.M. Robinson, F.-C. Tsui, J.U. Espino, W.R. Hogan. Design of a national retail data monitor for public health surveillance. Journal of the American Medical Informatics Association 10/5 (Sept/Oct), 409-418, 2003.

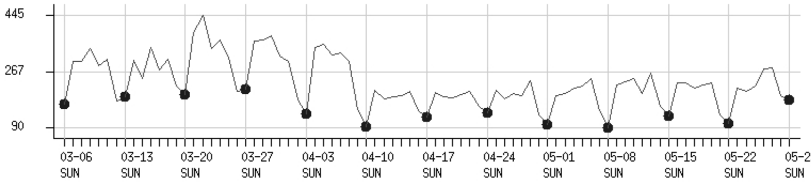


Figure 1. Weekly trend in Baby/Child electrolyte sales

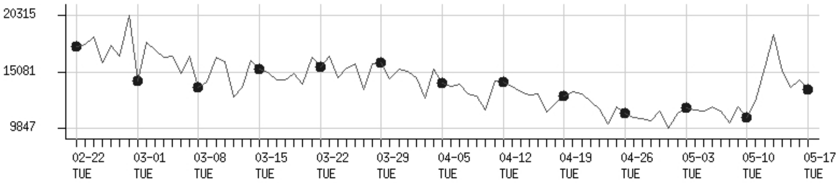
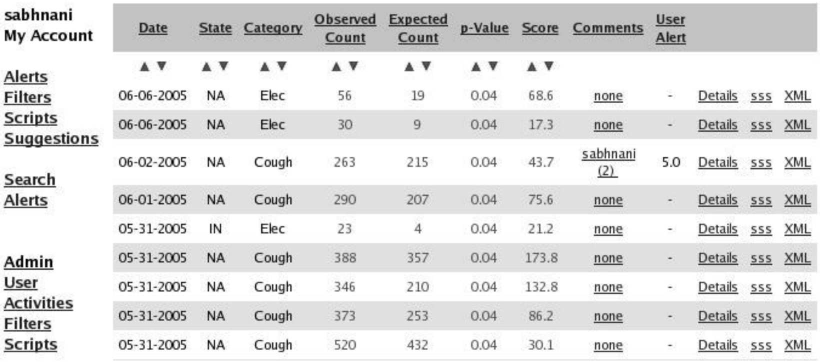
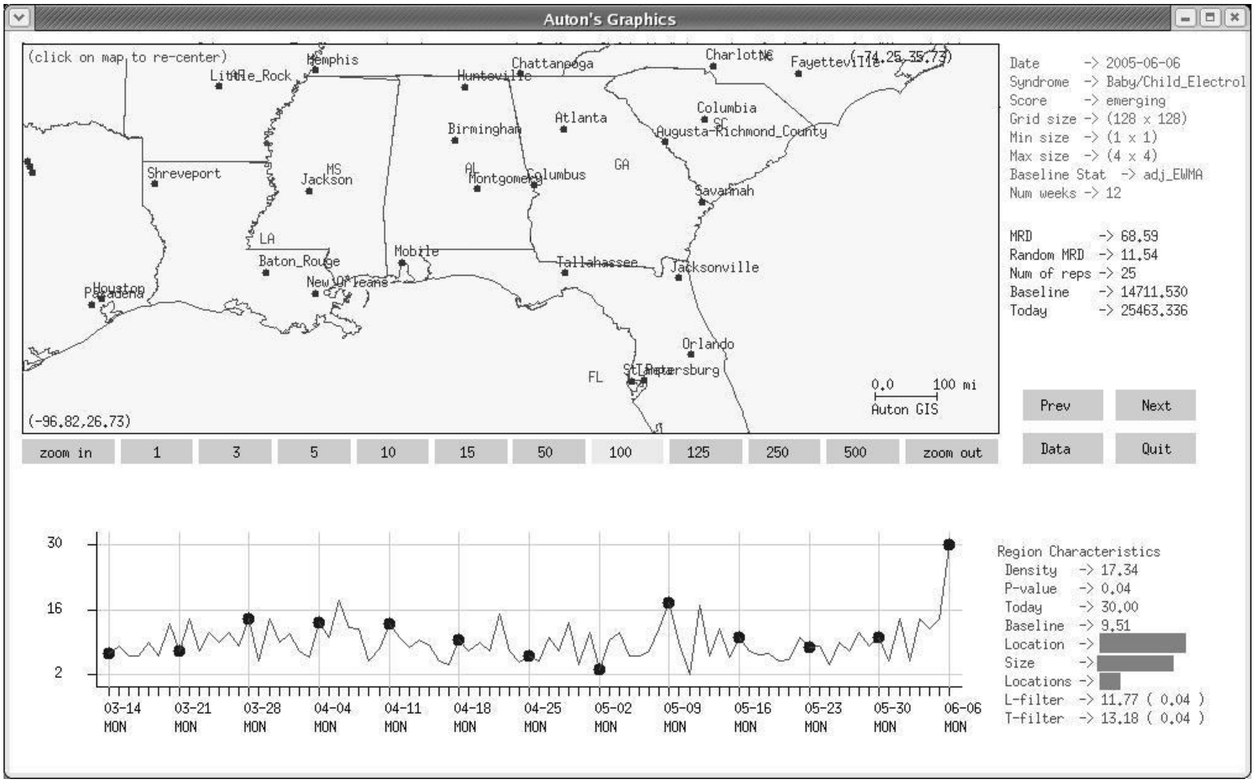


Figure 2. Seasonal trend in Cough/Cold sales



Date	State	Category	Observed Count	Expected Count	p-Value	Score	Comments	User Alert			
▲ ▼	▲ ▼	▲ ▼	▲ ▼	▲ ▼	▲ ▼	▲ ▼					
06-06-2005	NA	Elec	56	19	0.04	68.6	none	-	Details	sss	XML
06-06-2005	NA	Elec	30	9	0.04	17.3	none	-	Details	sss	XML
06-02-2005	NA	Cough	263	215	0.04	43.7	sabhnani (2)	5.0	Details	sss	XML
06-01-2005	NA	Cough	290	207	0.04	75.6	none	-	Details	sss	XML
05-31-2005	IN	Elec	23	4	0.04	21.2	none	-	Details	sss	XML
05-31-2005	NA	Cough	388	357	0.04	173.8	none	-	Details	sss	XML
05-31-2005	NA	Cough	346	210	0.04	132.8	none	-	Details	sss	XML
05-31-2005	NA	Cough	373	253	0.04	86.2	none	-	Details	sss	XML
05-31-2005	NA	Cough	520	432	0.04	30.1	none	-	Details	sss	XML

Figure 3. Screen shot of SSS user home page on the Web

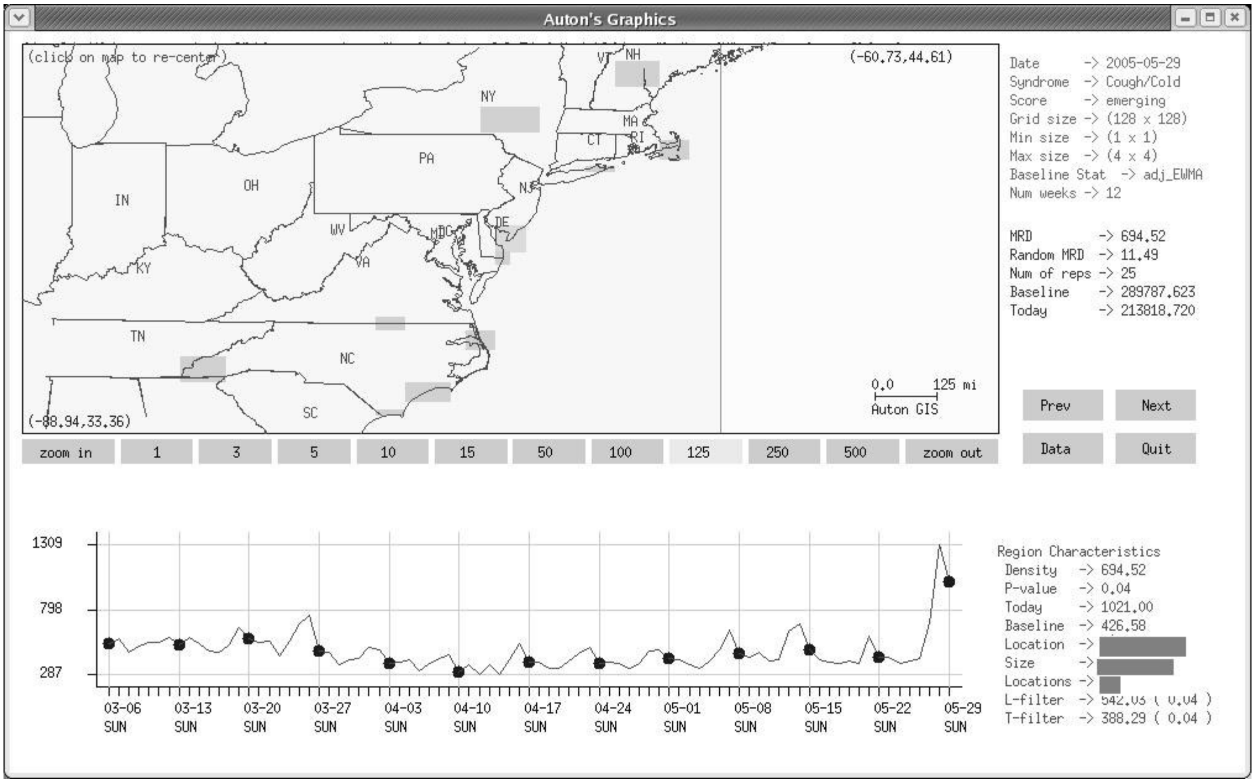


```
Date       -> 2005-06-06
Syndrome   -> Baby/Child_Electrol
Score      -> emerging
Grid size  -> (128 x 128)
Min size   -> (1 x 1)
Max size   -> (4 x 4)
Baseline Stat  -> adj_EWMA
Num weeks  -> 12

MRD          -> 68.59
Random MRD   -> 11.54
Num of reps  -> 25
Baseline     -> 14711.530
Today        -> 25463.336
```

```
Region Characteristics
 Density   -> 17.34
 P-value   -> 0.04
 Today     -> 30.00
 Baseline  -> 9.51
 Location  ->
 Size      ->
 Locations ->
 L-filter  -> 11.77 ( 0.04 )
 T-filter  -> 13.18 ( 0.04 )
```

Figure 4. Potential disease outbreak at the border of Alabama and Georgia



```
Date       -> 2005-05-29
Syndrome   -> Cough/Cold
Score      -> emerging
Grid size  -> (128 x 128)
Min size   -> (1 x 1)
Max size   -> (4 x 4)
Baseline Stat  -> adj_EWMA
Num weeks  -> 12

MRD          -> 694.52
Random MRD   -> 11.49
Num of reps  -> 25
Baseline     -> 289787.623
Today        -> 213818.720
```

```
Region Characteristics
 Density   -> 694.52
 P-value   -> 0.04
 Today     -> 1021.00
 Baseline  -> 426.58
 Location  ->
 Size      ->
 Locations ->
 L-filter  -> 542.03 ( 0.04 )
 T-filter  -> 388.29 ( 0.04 )
```

Figure 5. Long weekend trend showing the tourist spots in the country