Analyzing over-the-counter medication purchases for early detection of epidemics and bio-terrorism

Anna Goldenberg
CALD, Carnegie Mellon University
5000 Forbes Ave,
Pittsburgh, PA 15213
anya@cmu.edu

Abstract

Early detection of epidemics and bio-terrorism attacks is one of the most vital problems attempting to prevent deaths and provide for the opportunity of early treatment. There are many sources of data for analysis, though they are hard to get access to. However, there is very little real-time data that is readily available in substantial quantities at the first signs of an epidemic or a bio-terrorism attack. The main goal of this project is the investigation of the possibility of an early detection by analyzing trends of consumer purchases of over-the-counter medications. The idea is that by taking a large enough time frame of the regular over the counter medication purchases it might be possible to detect abnormalities early enough so that lives would be saved. One of the major parts of the research is the creation of a scalable high performance data retrieval system. It is important to ensure that the basic data access and data manipulation functions are performed as efficiently as possible on the dataset that could possibly be on the order of millions of records. Such a system may be created by inspection of the capabilities of commercially available database packages or by combining existing efficient data storage/retrieval methodologies. Another important part of the project is the analysis of the real-time data. Since we are interested in early detection or forecasting, time series prediction approaches are to be considered. It is expected that due to the particular nature of the problem, a combination of one or more techniques would have to be utilized. As a result of the project, it is desirable to have a system that would detect a signal in the regular over-the-counter medication purchases if the signal were present. This work might have a significant impact on the evolution of early detection and consequently on computer-based surveillance (CBS) systems, since the main approach being explored now is from the medical field, such reporting systems for clinicians, rather than evaluation of nonspecific syndrome data, such as over-the-counter medications.

1. Introduction

The problem of epidemics and bio-terrorism attacks is related to public health and therefore is one of the most vital but also controversial problems. The possibility of early detection of epidemics has been researched for centuries. There were many ideas and plans implemented in order to prevent or detect the epidemics as early as possible. However, even though vaccines to prevent the known severe deceases were developed, when the new strain of known or an unknown virus is introduced to the environment, the course of its evolution and spread is up to an early detection system. Unfortunately, early detection systems are still in the stage of development.

The notion of bio-terrorism attacks is rather recent but has attracted attention of the US government already. One of the key incidents that was a triggering point in persuading president Clinton to increase the budget to $146 millions annually (Stephenson, 1997; Associated Press 1999; Jacobs 1999), was the case of the Rajneeshee religious cult in The Dalles, Oregon. The cult planned to infect residents with *Salmonella* on election day to influence the results of county elections (Torok et al 1997). To practice for the attack, they contaminated salad bars at 10 restaurants with *S*. Typhimurium on several occasions before the election. A community-wide outbreak of salmonellosis resulted; at least 751 cases were documented in a county that typically reports fewer than five cases per year. Although bioterrorism was considered a possibility when the outbreak was being investigated by public health officials, it was considered unlikely. The source of the outbreak became known only when FBI investigated the cult for other criminal violations. A vial of *S.* Typhimurium identical to the outbreak strain was found in a clinical laboratory on the cult's compound, and members of the cult subsequently admitted to contaminating the salad bars and putting *Salmonella* into a city water supply tank. This incident, among other recent events, underscores the importance of improving preparedness at all levels.

The investigation of prevention systems is in the hands of the government officials. The development of such a system is a hard task to accomplish. A more feasible but not less desirable task is a creation of an early detection system. Though it seems that the task is simplified, it remains a non-trivial problem. Several important questions are to be answered: what are the sources that first notice the increase of the public health threat? Where is the information that could hint the first signals of sickness in a large group of people? Most people do not call emergency services immediately but start by taking pain killers, looking for symptoms on the web, calling friends, etc. By the time, the medical community has received a substantial amount of calls to get concerned - it may be too late.

Several ideas were discussed by a large group of CMU and PITT researchers. It was proposed to try to collect the information from the web and look for spikes in the hits to particular sites, another idea was to use the apparatus designed by BodyMedia (see www.bodymedia.com for more information) in order to track the body measurements to identify abnormalities. Unfortunately, the implementation of ideas requires substantial amounts of information that is hard to get in case of web sources or unavailable due to a small number of participants in the BodyMedia program. On the other hand, among other valuable sources is the large supermarket dataset that contains the information of over-the-counter medication purchases. It is possible that the data is noisy, but the potential of discovering the first signals of seasonal flu that in turn might be an indication of a bio-terrorism attack (see Goldman 2000) might be enormous.

By taking a large enough time frame of the regular over-the-counter medication purchases it might be possible to detect abnormalities early enough to prevent the epidemic from spreading or to provide the early treatment to the victims of the bio- attack. There are several time series predictions approaches that can be inspected in order to deal with the problem. Several possible methods will be discussed in the Approaches section. However, the analysis is not possible without the creation of a scalable, high performance data retrieval system. I would like to think of it as a stand-alone phase of the project. It is always important to ensure that the basic data access and data manipulation functions are performed as efficiently as possible especially since the dataset could possibly be on the order of millions of records. As it stands at the moment the commercially available tools lack one of the necessary qualities: either scalability, fast access or concurrency control is always an issue. Therefore, if fast, memory efficient algorithms are developed in order to tackle the problem at hand, they might become an important discovery in the area of databases and data structures.

The complications in recovering signal from noise are significant. First of all, people buy over the counter medications not only in cases of illness, but also to replenish their home reserves in case of future sickness. Also, sales and mail-in coupons might be a factor in increased amounts of purchases of the over-the-counter medications. It also needs to be acknowledged that the purchase series fluctuate depending on the day of the week and time of the year.

It is expected that this project will answer questions such as: can the beginning of an epidemic or a presence of an event of a bio-terrorism attack be detected using just the over-the counter medications, provided the data collected? Is it necessary to combine several sources to obtain a signal that would be strong enough? Also, there is intent to carry out substantial database-related work.

2. Prior Work

Sponsored by Defense Advanced Research Projects Agency, DARPA, there have arisen many initiatives for creating early detection systems, most of them however are considered with slightly different issues, such as developing a prototype database and Web site to give clinicians the opportunity to report suspicious trends of possible bioterrorist events; assessing the current capacity of hospitals and health systems to respond to a bioterrorist attack; or evaluate and improve linkages between the medical care, public health, and emergency preparedness systems to improve detection of and response to a bioterrorist event. [1] Since the “bio-terrorism awareness” project has been open relatively recently and most of the initiatives do not overlap, there is not much history in developing systems monitoring real-time data in order to prevent bioterrorist attacks and epidemics.

---

[1] This information is coming from the AHRQ meeting on Analysis of Detection Problems and is highly confidential. Please do not copy or distribute.

Most of the analysis performed on the supermarket data has also had a slightly different purpose. The research, often sponsored by the private corporations owning the particular supermarket chain, has concentrated on development of pricing strategies and analyzing the pricing sensitivity (see Boatwright, McCulloch and Rossi, 1999; Montgomery 1997, etc.)

Even though the subject of evaluating the trends in over-the-counter medications purchases to detect epidemics or bio- attacks has not been discussed in the literature, there are good overview papers that provide insight into database handling techniques as well as time series prediction approaches.

In the past decade, advances in speed of commodity CPUs have far out-paced advances in memory latency. Main-memory access is therefore increasingly a performance bottleneck for many computer applications, including database systems. The insights gained previously are translated into guidelines for database architecture; in terms of both data structures and algorithms. Manegold, Boncz and Kersten discuss how vertically fragmented data structures optimize cache performance on sequential data access. They then focus on equi-join, typically a random-access operation, and introduce radix algorithms for partitioned hash-join. The technique seems to be promising though there are some known drawbacks to using hash implementations and hash-joins. Some interesting guidelines are provided in the book by Han and Kamber, 2000.

The time series prediction analysis has a long history starting before 1920s, when the forecasting was done by simply extrapolating the series through a global fit in the time domain. With the Yule's invention of autoregressive technique in 1927 and for almost fifty years following his discovery the paradigm was that of linear models driven by noise (Weigend and Gershenfeld, 1993). However, for our case the linear modeling is most likely inadequate, since the process of purchasing tends to be highly nonlinear. Starting 1980s more complicated mixtures of linear models as well as non-linear approaches were explored. Some of the non-linear approaches will be discussed in the next section.

3. Approach

As it has been mentioned the project consists of two major parts. A lot of attention has to be paid to data representation and data structuring. Since it is expected that substantial analysis would be performed on the data, it is very important that data access and data processing are fast and reliable. The second and major part of the project is analysis of data and inference. Is it possible to recover signal from noise? Is there enough data collected? All of these questions and more are to be answered during the second phase of the project, the analysis phase.

3.1 The first phase – Data Management

There are many possibilities to structure the data in order to provide for fast access and quick searching. It is important to understand however, that the amounts of data are not trivial. Each week of data is stored on magnetic tape and consists of approximately 40 million records with information about the customers, their purchases and items they have purchased. This data represents a larger bulk than needed for the particular project, however it is not trivial to create the particular extraction. Several attempts were made to organize the data and populate the database. The Universal Database (UDB) by IBM was offered as a choice for storing and managing the data, however as it stands at the moment eight hours are needed to transfer the data from the tape to the database and another two hours to be able to query the data, during that time the database remains locked. Keeping in mind that the process takes ten hours per week of data and that the week time frame is too small to pick up any reliable signal, it is clear that further efforts need to be made to improve the performance of data retrieval and access. It is possible that in collaboration with IBM the usability of the database can be greatly improved, however taking into account the size of the data and the growth of the database, it may become necessary to use flat files applying memory efficient scalable data structures, such as AD-trees (Komarek and Moore, 2000) or KD Trees(Moore, 91), for data storage. It is possible that neither of the suggested techniques achieves the desirable performance, in which case new algorithms would have to be devised for the task at hand.

There are many tools and commercial packages available for data mining, however most of them were tested on the simulated datasets and the performance figures were estimated using theoretical derivations rather than by testing the tools on the real large scale datasets, presented in this problem. The experience shows that the time it takes to process the data available is prohibitive under any circumstances. This presents another challenging task to the problem. There needs to be developed a user friendly tool that allows for at least the simple data extraction and data manipulation operations to be performed so that the further analysis can be conducted.

3.2 The second phase – Data Analysis

It is possible to see from the above that the data storage, representation and manipulation is an important and challenging part of the project. The main task of the project though is to analyze the data for presence of a signal of increased sale of the over the counter medications that would hopefully correspond to the beginning of an epidemic or a possible occurrence of the bio-terrorist attack.

There are several stages to the analysis process as well. First of all, it would be essential to discover whether all the significant data is being collected in order to make inferences about the trends in medication purchases. The condition to check is whether enough data is collected time-wise, i.e. we would like to detect an abnormality in the data flow, therefore we need to have a large enough timeframe to be able to establish the regular flow and enough data containing as much information about abnormalities as possible. If it is realistic to assume that there is enough data to be able to detect an epidemic, what should we be looking for? Certainly, the aggregate amounts of analgesics sold per day might form an interesting curve to observe and the picks of the curve might actually give an idea about some irregularities in the day's sale, that may in turn be a hint for an epidemic, but is it really so? How can we "early detect" the epidemic next time if we are waiting for the pick to occur, where pick in this case might coincide with the pick of an epidemic. This kind of detection can hardly be deemed as an "early" one.

The problem at hand can be thought of as the time-series analysis prediction. It is not expected to predict the next epidemic though; it would suffice to make the detection by combining the predictions and tracking the data online. The hope is that the approximate prediction would help in faster detection.

A number of techniques are available as a starting point. Theoretically it is possible for the data to be linear or non-linear, however it has been known that the real-life processes can very rarely be well approximated by linear models. One of the approaches is to try to reconstruct the state space of a dynamical system. It is perceived that the system of the series of purchases can be described by a number of "state variables" that determine what the values of the time series should be (Chatfield, 1988), for example the state of epidemic would be one state, the state of store-sale would be another state (the states defined as such would have to be able to overlap in time), etc. We may proceed by guessing how close conditions at various times in the past are to present conditions and then appropriately averaging near matches to make forecasts. Here, closeness represents the distance in state space and the guesses are implemented by conditional probability density function for location in state space, given observations (Weigend and Gershenfeld, 1993). It is then possible to apply Hidden Markov Models (HMMs), where the probability distributions can be adjusted by the Baum-Welch, or forward-backward, algorithm to maximize the likelihood of a given set of training data.

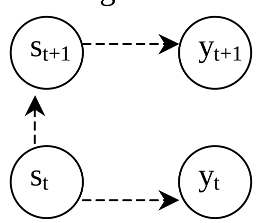

Figure 1. HMM (see (Weigend and Gershenfeld, 1993) for more information)

Another approach making use of the state space interpretation of data is the delay coordinate embedding (see Takens, 1981; and Sauer, Yorke, and Casdagli, 1991). The method is based on the assumption that short-term predictions of deterministic time series is possible, even if the system producing the time series is chaotic (Echmann and Ruelle, 1985), as it seems to be in the given situation. This method provides an elegant solution for reconstructing the current state from the time series. It assumes a theoretical state a and then uses the available data to find a vector b in a Euclidian space R^m , with which to replace the theoretical state. This method though appealing, has many unresolved issues. The method is highly dependent on the techniques used to estimate the prediction function $P_t(a)$, i.e. depending on the choice of a function used to estimate the future state t given the present state a will come up with significantly different predictions.

An alternative powerful tool that can be applied to the over the counter medication purchases series in attempt of early detection of epidemics or bio- attacks is a connectionist network with internal delay lines or a neural network architecture modeling synapses as FIR (Finite Impulse Response) linear filters. The internal dynamics is brought to the basic neural network by replacing each static synaptic weight by a FIR filter:

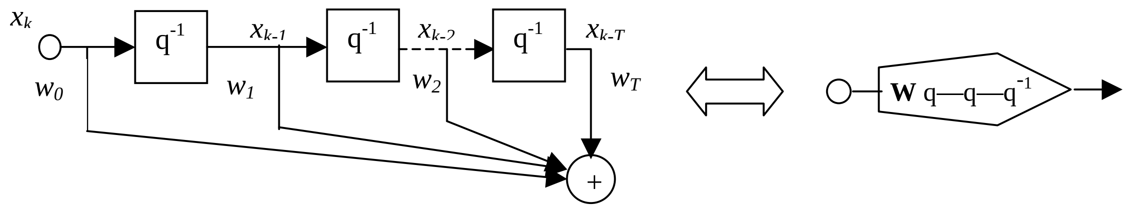

Figure 2. FIR filter model, where q^{-1} represents a unit delay operator $x_{k-1} = q^{-1}x_k$

The computational complexity and the time constraints are significant drawbacks of applying this method. It however is a powerful technique that should definitely be inspected.

It is important to mention fractal methodology as a possible tool in the prediction of time series analysis in over the counter medications. The important idea behind the approach is self-similarity of the series. It is not unfair to assume that the epidemics, in-store sales are self-similar patterns. It is then possible to estimate the fractal dimension of the flue epidemic curve and then search for a time frame of similar size for a series with a similar fractal-dimension. This approach also has its drawbacks. When talking about the early detection systems, it is not enough to have a good analysis tool though it would be of definite value. It is important to save time and make the detection as early as possible, thus the fractal dimensionality though a valid tool might have to be used in combination with others to make early predictions when evaluating the real-time data.

3.3 Complications

At this moment, it is hard to predict how much information the data may contain. Though the tools described above are all valid, the problem might arise from the data itself. What if there is just not enough signal in the noise? Perhaps, the data by its nature does not contain enough information to be a valid source of epidemic detection by itself. In this case, there are several other data sources, such as web click-stream data from certain medical sites or may be BioMedia data that might scale up by the time the project is in progress, that would in combination with the supermarket data create a strong enough signal in order to detect patterns of arising epidemics and/or bio-terrorism attacks.

3.4 Approach Validation Techniques

From the description of the problem and the approach above, it can be seen that the simplest way to evaluate the system is make inferences about an event of arising epidemic and then wait to see if it happens. This approach however, might be too expensive in terms of well-being of many people. Another possibility is cross-validation. If there are multiple occurrences of epidemics in the dataset, it is possible to withhold some data, training on one sequence and testing on the other. It would be ideal to have an separate unrelated source of data so the system could be tested against it, however such data is hard to obtain and is not available at the moment.

It is anticipated that the methods described above, one or in combination with each other or perhaps some alternative methods, may all together yield an analysis mechanism applicable to the real-time data. It is expected that by identifying techniques for fast access and easy manipulation of the data developed in the first phase of the project and combining them with the detection/prediction tool designed in the second phase, it will be possible to help in early detection of epidemics or bio-attacks by incorporating the system into the regular every day monitoring of public health through in-store purchases of over the counter medications.

5. Work Plan

As described previously, the project consists of two phases. As the research proceeds, each of them might become a project by itself, since they have to be done in sequence and each represent a non-trivial task.

It is proposed here that the first part of the project is completed within three weeks starting February 9th:

- Feb 9th – Feb 18th: analysis of the capabilities of the UDB – trying to improve the performance of the existing database. Possible restructuring.
- Feb 19th – Feb 26th: exploration of other data structures, such as custom AD-trees, KD-trees, etc.

The second part is to start on Monday, February 26th:

- Feb 26th-March 3rd: explore what can be inferred from data by trying to weigh different parameters by different levels of significance
- March 4th-March 20th: apply various techniques discussed above to create a prototype of a detection system

6. Summary

This project proposes an alternative to conventional Computer Based Surveillance Systems (CBS), by concentrating more on the non-specific syndrome data such as over-the-counter medication. The idea is based on the time series prediction analysis of the series of medication sales in attempt to recover the signal of epidemics or possibly the bio-terrorist attacks. The advanced database techniques will be applied in order to create a flexible, scalable data access/retrieval system, providing for easier data manipulation in the analysis stage. It is important to notice that there is a possibility that the signal in the data is not strong enough. As a result of the project, it is desirable to have a system that would reliably detect a signal of abnormality in the regular over-the-counter medication purchases if the signal were present. This work might therefore have a significant impact on the evolution of early detection systems, due to the uniqueness of its nature.

References

1. Torok T.J., Tauxe R.V., Wise R.P., Livengood J.R., Sokolow R., Mauvais S., et al. A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. *JAMA*;278:389-95, 1997.

2. Montgomery, A. Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data. 16, 4, 315-337, 1997

3. Weigend, A.S. and Gershenfeld, N.A., "Time Series Prediction: Forecasting the Future and Understanding the Past", Addison Wesley, 1993.

4. Markus, C., Gupta, S., Porter, J. C., Staelin, R., and Wittink, D.R. Using Market-Level Data to Understand Promotional Effects in a Nonlinear Model, 34 (August), 322-334

5. Boatwright P., McCulloch R., and Rossi P.E. Account Level Modeling for Trade Promotion: An Application of a Constrained Parameter Hierarchical Model, 1999

6. Han, J. and Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000

7. Chatfield, C. What is the best Method in Forecasting? J. Appl. Stat. 15: 19-38,1988

8. Takens, F. Detecting Strange Attractors in Turbulence. In Dynamical Systems and Turbulence, edited by D.A. Rand and L.-S. Young. Lecture Notes in Mathematics, Vol. 898, 336-381. Warwick, 1980.

9. Sauer, T., J.A. Yorke, and M. Casdagli. Embedology. J.Stat. Phys. 65(3/4): 579-616, 1991.

10. Eckmann, J.P,. S.O Kamphorst, and D. Ruelle. Recurrence Plots of Dynamical Systems. Europhysiscs Letters 4: 973, 1987.

11. Goldman B, Inhalation Anthrax in Cecil Textbook of Medicine. Saunders Company. New York. 2000

12. P. Gassner, G. Lohman, B. Schiefer and Y. Wang. Query Optimization in the IBM DB2 Family. In IEEE Data Engineering Bulletin, 16:4-18, 1993.

13. Agrawal, R. and K. Shim (1995). Developing tightly-coupled applications on IBM DB2/CS relational database system: Methodology and experience. Research Report RJ 10005(89094), IBM Corporation

14. C. Faloutsos Searching Multimedia Databases by Content, Kluwer Academic Press, 1996

15. Paul Komarek and Andrew Moore, A Dynamic Adaptation of AD-trees for Efficient Machine Learning on Large Data Sets, International Conference on Machine Learning, 2000.

16. Stephenson J. Pentagon-funded research takes aim at agents of biological warfare. *JAMA*;278:373-375, 1997

17. Associated Press. Lab meets with residents over bacteria release. New York, *NY: Associated Press*; July 9, 1999.

18. Jacobs P. Attack of the killer microbes. The Pentagon is funding research into novel ways to thwart bioterrorists. Advances could pay off commercially as well. *Los Angeles Times*. August 19, 1999.

19. A. W. Moore, An introductory tutorial on kd-trees. Extract from A. W. Moore's Phd. thesis: Efficient Memory-based Learning for Robot Control, Computer Laboratory, University of Cambridge, Technical Report No. 209, 1991

20. Rülleke, T.; Fuhr, N. Probabilistic Reasoning for Large Scale Databases. In: Datenbanksysteme in Büro, Technik und Wissenschaft, pages 118-132. Springer, Berlin et al., 1997

21. P. Ciaccia, M. Patella, and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In Proceedings of the International Conference on Very Large Databases (VLDB), Athens, Greece, 1997

22. J. Pastor, D. McKay, and T. Finin. View--concepts: Knowledge--based access to databases. In Proceedings of the First International Conference on Information and Knowledge Management, October 1992.

23. Ramez Elmasri, Gene T. J. Wuu, and Yeong-Joon Kim. The time index: An access structure for temporal data. In Proceedings of the 16th International Conference on Very Large Databases, pages 1--12, August 1990

24. S. Manegold P. A. Boncz and M. L. Kersten. Database architecture optimized for the new bottleneck: Memory access. In 25nd Intl. Conf. Very Large Databases, September 1999.