Information Systems 48 (2015) 151–163

Contents lists available at ScienceDirect

Information Systems

journal homepage: www.elsevier.com/locate/infosys

An analytics appliance for identifying (near) optimal over-the-counter medicine products as health indicators for influenza surveillance

Ruhsary Rexit [a,b,*], Fuchiang (Rich) Tsui [a,*], Jeremy Espino [a], Panos K. Chrysanthis [b,*], Sahawut Wesaratchakit [a], Ye Ye [a]

[a] The RODS Laboratory, Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA 15260, United States
[b] The ADMT Laboratory, Department of Computer Science, University of Pittsburgh, Pittsburgh, PA 15260, United States

ARTICLE INFO

Available online 10 June 2014

Keywords:
OTC analytics appliance
Distributed search
Outbreak detection
Time series analysis
Syndromic surveillance

ABSTRACT

In the era of "Big Data", a challenge is how to optimize our use of huge volumes of data. In this paper, we address this challenge in the context of a public health surveillance system which identifies disease outbreaks using individual and population health indicators. Our goal is to automate and improve the accuracy of the selection process of the health indicators, a process which is data-intensive and computationally expensive. The health indicators selection process traditionally has been carried out manually by public health experts in collaboration with health data providers. In particular, we present an approach for identifying sets of over-the-counter (OTC) medicine products whose aggregate sales correlate optimally with aggregate counts of emergency department (ED) visits. Towards this goal, we propose an OTC Analytics Appliance which utilizes a distributed search engine to efficiently generate time series of time-stamped records and supports "plug-and-play" search and correlation functionalities. Using the OTC Analytics Appliance with the Pearson correlation coefficient function, we evaluate Brute-force search, Greedy search, and Knapsack search for their ability to select the optimal or suboptimal set of OTC products automatically. Our results show that greedy search is the most preferable, producing a set of OTC products whose sales that correlate optimally or near optimally to ED visits, while achieving acceptable search times with large datasets. Also, our evaluations show that our approach using the greedy search can be potentially used to efficiently identify different optimal OTC medicine products for detection of different types of disease outbreaks.

1. Introduction

An outbreak detection system is a public health surveillance system used for identifying increases in the incidence rate of a disease. A Syndromic Surveillance system is a type of outbreak detection system that monitors the health status of a community and identifies outbreaks using individual and population health indicators. Those health indicators are available before a confirmed diagnoses or laboratory confirmation [1]. Syndromic Surveillance systems have used various data sources as health indicators, including over-the-counter (OTC) medication sales, emergency department (ED) chief complaints, school absenteeism data, and web search queries [2–4]. Among these, ED data generally serves as the core data source of many Syndromic Surveillance systems such as BioSense [5] and RODS [2]. Researchers have shown that common outbreaks can be

* Corresponding authors.

http://dx.doi.org/10.1016/j.is.2014.05.008
0306-4379/

detected 1–2 weeks earlier with ED data than through conventional disease reporting methods [6].

A common methodology employed in Syndromic Surveillance systems is to aggregate health-related temporal events into time series that are analyzed algorithmically for the detection of outliers. The intuition of using OTC medication sales is that sick individuals typically purchase some OTC medications to treat themselves before seeing a doctor. For example, in an Syndromic Surveillance system using OTC medication sales as an indicator to detect influenza outbreaks, the epidemiologist would analyze a time series of the daily sales of all cough syrup, thermometers, and fever reducers in a specific geographic region. If daily sales of these products exceed some threshold (e.g., three times the standard deviation from a baseline value), that could indicate a disease outbreak.

The effectiveness of a Syndromic Surveillance system depends on three factors:

1. The availability of health related data from providers e. g, hospitals, food and drug retail industry.
2. The selection of good health indicators (such as specific medications sold) to be used for the detection of a disease outbreak by the Syndromic Surveillance systems.
3. The ability of the Syndromic Surveillance system to provide query processing and data analyses *on the fly*, as the data arrives at the system.

Both the bootstrapping, i.e., the selection of health indicators, and the outbreak detection activities of a Syndromic Surveillance system during its deployment involve data filtering and spatio-temporal aggregation over time series. The key difference is that in the former case of the selection of health indicators the query processing and analysis is carried out on historical data as opposed to the latter case of detection which is carried out on current data.

Despite this difference, the increasing volume of monitored OTC product sales in United States is a challenge for Syndromic Surveillance systems, we must be able to identify an optimal set of health indicators for a given syndrome as well as meet the near-real-time requirements of detection.

Initially, we employed and explored data warehouses and externally-implemented continuous queries, but we soon discovered that the performance of the datawarehouse- based online analytical processing (OLAP) was not able to support either the selection process or the detection process. The data warehouse approach required both large amounts of storage for storing the fact tables and pre-computed statistics and also incurred large overhead in first storing and then retrieving the data for analysis [7, 8]. For this reason we subsequently explored the use of (1) a data stream management system which efficiently execute continuous queries before storing the data [9–11], to support the detection process and (2) a distributed query processing system where filtering and aggregation take place over collaborating computers, possibly in the cloud, to support the selection process which often requires multi-year worth of baseline data. In this paper, we present the result of our exploration of using a distributed search engine to implement the selection process for identifying the optimal set of health indicators.

Specifically, our work in this paper was motivated by three observations: (1) traditionally, the selection process has been performed manually by public health experts, (2) it was limited by the amount of data used, and (3) traditionally, the analysis was centered in the measuring of the relationship of a manually selected set of health indicators to some survey, such as of sick individuals or hospital visits. These observations formulated our hypothesis that *if we want to accelerate the selection process and make it more accurate, then we need an efficient solution for processing large volume of aggregated data and time series that automatically identify the optimal set of health indicators*.

To support our hypothesis, we developed an *OTC Analytics Appliance* that efficiently generate time series of time-stamped records such as unit sales of certain OTC products and used it to compare different search algorithms to identify a set of thermometer products (such as strip or digital, oral or forehead for babies or adult thermometers) whose sales over time optimally, or close to optimally, correlates with ED visits for symptoms (such as fever) consistent with Constitutional syndrome.

Contributions: The two key contributions of this paper is as follows:

- The development of an OTC Analytics Appliance, which utilizes a distributed search engine, called ElasticSearch [12], to efficiently generate time series of time-stamped records. The OTC Analytics Appliance provides an Optimal OTC Identifier module with “plug-and-play” search and correlation functionalities and a GUI to display time series graphs.
- An evaluation of three search algorithms, *brute-force search*, *greedy search*, and *dynamic programming* (knapsack search) for their ability to select the minimum set of OTC products automatically. Our results using the Pearson correlation coefficient function show that greedy search is competitive to the brute-force search, producing a set of OTC products whose sales optimally correlate to ED visits, while at the same time maintaining scalability with large datasets. The knapsack search exhibits the worst performance. Also, our evaluations show that our approach using the greedy search can be used to efficiently identify different optimal OTC medicine products for detection of different types of disease outbreaks.

Roadmap: Section 2 introduces the OTC Analytics Appliance. Section 3 presents our experimental datasets and methods. Section 4 presents the experimental results for the optimality of the three search algorithms using data collected during one year period. Section 5 evaluates the robustness of the three search algorithms whereas Section 6 presents their scalability evaluation with an extended dataset of four years. Section 7 briefly reviews related studies and Section 8 concludes with future work.

2. System architecture

This section gives an overview of OTC Analytics Appliance, then explains each module of the system such as