Global Epidemiological Outbreak Surveillance System Architecture

Ricardo Jorge Santos[1] and Jorge Bernardino[2, 1]
[1]CISUC – Centre of Informatics and Systems of the University of Coimbra – University of Coimbra
[2]ISEC – Engineering Institute of Coimbra – Polytechnic Institute of Coimbra
PORTUGAL
lionsoftware.ricardo@gmail.com, jorge@isec.pt

Abstract

Diseases such as avian influenza, severe acute respiratory syndrome (SARS) and Creutzfeldt-Jacob syndrome represent a new era of biological threats. Nowadays, these hazards breed, mutate and evolve at tremendous speed. Furthermore, they may spread out at the same speed as which we travel. This reveals an urgent need for an agent capable of dealing with such threats. Data warehouses are databases which provide decision support by on-line analytical processing (OLAP) techniques. We present the architecture for an effective information system infrastructure enabling the prediction and near real-time detection of disease outbreaks, using knowledge extraction algorithms to explore a symptoms/diseases data warehouse in a continuous and active form. To collect such data, we take advantage of the Internet and features existing in today's common communication devices such as personal computers, portable digital assistants and cellular phones. We present a case-simulation based on a small country, showing the system can detect an outbreak within hours or even minutes after its physical occurrence, alerting health decision makers and providing quick interaction and feedback between all users. The architecture is also functionally independent from its geographical dimension.

1. Introduction

A data warehouse (DW) provides information for analytical processing, decision making support and data mining tools. A suitable data model is the core of representing part of the real world in the context of a database. Although many modeling techniques expressed in extended multidimensional data models were proposed in the recent past [5], many major issues such as information system architectures for specific health issues are not properly reflected.

Diseases such as avian influenza, SARS and the Creutzfeldt-Jacob syndrome represent a new era of biological threats. New stripes of viruses and bacterias are becoming increasingly aggressive and rapidly adapting to resist vaccines and medication. The speed at which these diseases are mutating and evolving, combined with the fact that they may spreadout at the same rate as people and animals travel, greatens the risk for a major epidemic or pandemic outbreak. It is therefore crucial to detect when a potential outburst might by taking place in order to contain it as quickly as possible and minimize damage it may cause.

Our architecture fulfils that need, using knowledge extraction algorithms to explore a symptoms/disease DW, looking for patterns of symptoms to predict the occurrence of a potential outbreak. We also present an experimental evaluation using a case-simulation for a small country.

The rest of this paper is organized as follows. In section 2, we refer issues and existing solutions in epidemics and health information systems. In sections 3, 4 and 5, we respectively present our architecture, its database and the main algorithms and methods for outbreak prediction and detection. In section 6 a simulation of the system working for a small country such as Portugal is presented and the final section contains concluding remarks and future work.

2. Background and related work

Accessing the Internet today, we can find several institutional and enterprise web portals which provide trustworthy health information (including epidemic and pandemic) such as in [1] by the Aberdeen Group, [7] by Great Britain's NHS, the World Health Organization [10]. We can also use web applications to perform a risk analysis on contagious diseases which can be disseminated through animal contact [9]. The work in [4] refers the importance of mathematical models given historical disease data as a mean of predicting and evaluating forms of action in certain situations. We can also use the Internet for reporting diseases to adequate health services, like what is done by the United States' Centre for Disease Control in what they refer to as "communicable diseases". However, with new emerging diseases, using historical data based contention plans will not be an efficient way to handle the problem, as shown in [8]. Innovative solutions have emerged based on telecommunication and informatics technology, such as the EMPHIS Project [2], following the perspective and vision of the future presented in [6] by Great Britain's NHS. The architecture presented takes the next step, combining database, knowledge extraction and telecommunication technologies to aid global health in rapidly predicting and/or detecting the occurrence of epidemic outbreaks, which is vital for minimizing losses and containing potential hazards.

3. The surveillance system's architecture

The technological evolution in telecommunications and portable computerized devices makes it possible today to have real-time information availability, practically without geographical dependencies. Taking advantage of an agent with the highest level of availability such as the Internet, our architecture provides the infrastructure for collecting data of occurring symptoms and diseases, and points examples on how to effectively and efficiently process this data to discover symptom and disease associations. This is done achieved by inserting patient symptoms and diseases data in a web server database, which collects all information in a given geographical region and ships it to a DW located in a health decision centre. If the number of discovered cases within that region is considered relevant as a possible epidemic indicator, health decision makers and medical staff are immediately alerted.

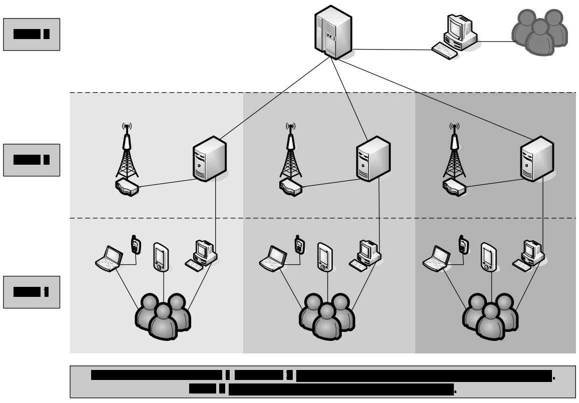

Figure 1. The surveillance system's architecture

The architecture has 3 bottom-up tiers or levels, as seen in Figure 1. Symptom/disease data is uploaded by medical staff using personal devices with internet access, such as mobile phones, PDAs or common personal computers, getting stored in the second tier web servers. Each web server has the database and software applications needed to support the first tier requested services. The decision making server in the last tier holds a DW processing non-stop knowledge extraction algorithms finding disease record counts and symptoms/disease patterns in a defined geographical area. If a relevant number of suspicious patterns of symptoms or confirmed occurrences of diseases are detected, health decision makers and medical staff are immediately alerted. Figure 1 represents an example of an implementation covering three defined geographical areas.

A major advantage in our proposal is that once the disease/symptom data is recorded, the detection process is much faster than bureaucratic processes used today. Nowadays, when a major disease is observed, medical staff fill in paperwork reporting those cases to entities such as the CDC in the United States or the NHS in Great Britain. These entities process and analyze the amount of cases received from each region and decide if that amount should be considered relevant. These processes usually take days, or, at least, many hours. Furthermore, if a "minor" disease is observed, such as a simple flu, for instance, it is not considered as relevant to report. Although it may be a "minor" disease, if it were to occur in a considerable amount of cases within the same region, it could become an important issue. With our system, this would be almost immediately detected and alerted; in the traditional existing processes it would not be detected, or, in the best case, would be noticed only after some time. For each medical staff disease or symptom input, they may not even physically know, see or even be in contact with each other, but their medical records will be matched almost in a real-time manner, detecting the possibility of an epidemic occurrence.

Each second tier web server must contain the following components in order to insure the systems interaction and functionality: a) a data mart containing the database structure and all supporting data for the geographical region and population it serves; b) a web interface for first tier users to input data and to promote interaction between third tier users (health decision makers) and first tier users (medical staff); c) a software application available to first tier users for downloading, which allows working offline the Internet and capable of uploading that data to the second tier web servers whenever requested. This would allow medical staff to work at any location without Internet access; d) a software server component responsible for shipping the collected data to update the third tier DW server.

4. The surveillance system's database

Today, most database systems offer features that go beyond management of static data and most information systems are powered by a database. The job of a database is to store data and answer queries. By contrast, the job of an information system is to provide a service, which are semantic entities entailing considerations that span the life cycle of the larger system [3]. Traditionally, database systems have been passive, storing and retrieving data in direct response to user requests without initiating any operations on their own. As the scale and complexity of data management increased, interest has grown in bringing active behaviour into databases, allowing them to respond independently to data-related events. Therefore, given the usage we wish to provide our database, we can look at it as an active database as discussed in [3], for it will be continuously querying and analyzing data and reporting it to the users makers involved in an interactive form.

The database holds patient symptoms and disease data records, including both humans and animals. Based upon the characterization of these entities and their attributes, we propose in Figure 2 the partial DW schema supporting human disease outbreak detection. The schema for outbreak prediction is similar and given by adding tables relating to symptom data. The schema for animal disease outbreak detection and prediction are similar to the human schema, linking each animal with the human to which it belongs.

The key tables for epidemic discovery are the fact tables: *Humans_Symptoms*, *Humans_Diseases*, *Animals_Symptoms* and

Animals_Diseases. The dimensional tables also play important roles: a) *Processes* represents each health appointment related to a human or animal symptom/disease; b) *Locations* indicates where it is happening; c) *Humans* and *Animals* to who it is happening; d) *Physicians* what medical staff is involved; e) *Institutions* and *Institutions_Physicians* where physicians can be reached; f) *Diseases* allows defining how many cases are considered relevant for an epidemic alert, for each disease, in absolute number or in % of the amount of population within a location; g) *Diseases_Symptoms* indicate possible symptom sets for each disease.

Due to space constraints, we will only explain attributes *LD_Relevance* and *LD_MinValue*, which play a vital role. The first defines which type of measure should be used for estimating the number of cases considered relevant for alerting a possible outbreak, while the second defines that number of cases for each geographical region. For example, if in a certain region 10 simultaneous cases is considered a potential outbreak of disease *X*, then *LD_MinValue* would be *10* and *LD_Relevance* would be *"Absolute"* (meaning that the value in *LD_MinValue* is an absolute value and does not depend on the population count of any geographical area), in the record where *LD_Disease_ID* is *X*. On the other hand, if the same disease *X* is considered as a potential outbreak when the simultaneous number of recorded cases represents 10% of the population, *LD_MinValue* would also be *10* but *LD_Relevance* would be *"Percent"*.

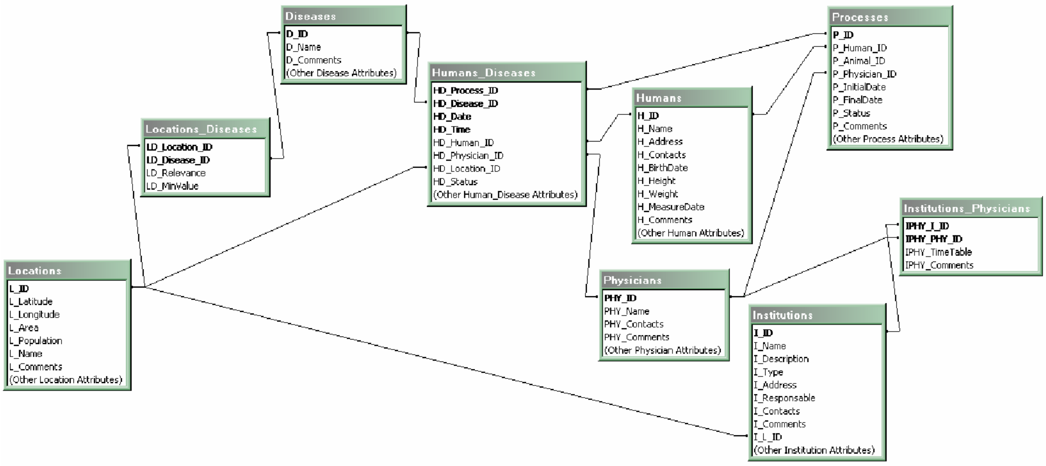

Figure 2. Partial database schema for human disease outbreak detection

5. Outbreak discovery

5.1. Detecting disease outbreaks

When looking for possible disease outbreaks, we count the number of simultaneous active disease records in each geographical area (location) and verify if that number is significant by comparing it with the number of cases which would be considered minimal for a possible disease outbreak in the location. For instance, suppose that for location *Y*, 9 cases of disease *X* is the amount of simultaneous cases from where there should be a possible outbreak alert. If there are 10 or more records of that same disease classified as active in that same region, the system should immediately alert health decision makers. This calls for an algorithm which will depend on disease identification attributes and counters.

Therefore, we may seek the *Humans_Diseases* and *Animals_Diseases* tables gathering all recorded active diseases for each process and grouping them per disease and geographical location. It is then possible to verify its significance by comparing it with the significant number of cases defined for each disease in the *Locations_Diseases* table for that location. An outbreak of disease D_i is defined if a significant number of active disease records for that disease D_i are registered within the same geographical area or location. The number of cases considered significant for each disease within each geographical area should be previously defined by health decision makers (referring to *LD_MinValue* and *LD_Relevance* attributes explained in previous section).

5.2. Predicting disease outbreaks

To predict an eminent outbreak we gather patterns of recent sets of symptoms per patient and relate them with diseases which have the same set of possible symptoms, accounting the amount of them. This calls for an algorithm depending on associating symptoms with possible diseases to which they might belong. To accomplish this, we use the set of recent registered symptoms in the each process recurring to *Humans_Symptoms* and *Animals_Symptoms* tables, to look for matching possible sets of symptoms of any disease by relating them with the *Diseases_Symptoms* table. This allows detecting a plausible outbreak ***before*** it happens, based on patient's recorded symptoms.

Consider a process for patient *P*, having a set of *n* symptoms, $PS_n = \{ PatientSymptom_1, PatientSymptom_2, \ldots, PatientSymptom_n \}$, and a disease *D* characterized by a set of *m* symptoms, $DS_m = \{ DiseaseSymptom_1, DiseaseSymptom_2, \ldots, DiseaseSymptom_m \}$. We have a possible existence of disease D in patient P if PS_n exists in DS_m: $DS_m \subset PS_n$

We have the existence of a possible outbreak of disease D if the counting of patients where $DS_m \subset PS_n$ is greater or equal to a number of cases considered significant as an indicator of an outbreak.

6. Experimental evaluation

To test the system, we built a disease/symptom dataset generator based on a nation such as Portugal. The country's area was divided into approximately 4.000 geographical locations, corresponding to 20 Km^2 each. To each location a random population of 1 to 5000 people was generated. One year of humans and animals processes, symptoms and diseases data was generated, distributing temporal data in an equally random form. Table 1 shows the records generated for each table of the system's database. This data was loaded into a unique server, a Pentium IV 2.8GHz with 1GB DDRAM, using Oracle 10g DBMS. After loading the DW its size was approximately 20GB. We used a Delphi 7 compiler to build an interface using SQL to implement the outbreak search algorithms. The record insertion rate for each fact table is presented in Table 2, with an average total of 6,4 disease/symptom related records per second.

After running the system simulating the online insertion of disease/symptom records it took approximately 14 minutes to achieve the listing of disease outbreaks and 45 minutes to obtain the listing of suspicious patterns of symptoms in the last 7 days, indicating the geographical locations which should be kept under surveillance. *This means we would have an outbreak detection in less than one hour from its recording occurrence!*

TABLE	NUMBER OF RECORDS
Humans	10.000.000
Animals	40.000.000
Physicians	100.000
Institutions	110.000
Institutions_Physicians	250.149
Diseases	3.597
Symptoms	277
Diseases_Symptoms	21.829
Locations	3.956
Locations_Diseases	14.229.732
Processes	40.000.000
Humans_Symptoms	89.996.523
Humans_Diseases	20.000.000
Animals_Symptoms	49.989.966
Animals_Diseases	9.999.329
DATABASE TOTAL:	274.705.358

Table 1. Record counting of the generated datasets

TABLE	INSERTION FREQUENCY
Processes	1 Record per 1,5768 seconds
Humans_Symptoms	1 Record per 0,3504 seconds
Humans_Diseases	1 Record per second
Animals_Symptoms	1 Record per 0,6308 seconds
Animals_Diseases	1 Record per 3,1538 seconds
Data Warehouse Total	≅ 6,4 records per second

Table 2. Disease/symptom data insertion frequency

7. Conclusions and future work

This paper presents a feasible solution for epidemic and pandemic outbreak near real-time prediction and detection, providing a solution for a major global public health issue. It allows continuous monitoring of possible disease epicentres, enabling early detection and even prediction of possible emerging epidemic and pandemic outbreaks. This is of major importance in maximizing contention and minimizing losses, namely human lives.

We have shown its functionality, recurring to a simulation of an entire country's symptoms and diseases data, being able to detect disease outbreak in less than one hour for a population of 10M people and 40M animals.

As future work, we intend to adjust the architecture for issues such as data processing in countries like China, with a massive population count.

8. References

[1] CRM Access Solution Finder > Health Directory, partners.knowledgestorm.com/search/tabkeyword/abdcrm/software/Health+/1/index.asp, *CRMA Solution Finder*, Aberdeen Group, 2004.

[2] EMPHIS Project, *Euro-Mediterranean Public Health Information System*, December 2002.

[3] D. Goldin, S. Srinivasa and V. Srikanti, "Active Databases as Information Systems", International Database Engineering & Applications Symposium, IDEAS 2004.

[4] S. Gouldie, "Preventing Cervical Cancer in Developing Nations", *Harvard Center for Risk Analysis*, July 2001.

[5] W. Hummer, W. Lehner, A. Bauer and L. Schlesinger, "A Decathlon in Multidimensional Modeling: Open Issues and Some Solutions", International Conference on Data Warehousing and Knowledge Discovery, DAWAK 2002.

[6] NHS Executive Report, "Information for Health – An Information Strategy for the Modern NHS 1998-2005", *Great Britain National Health System*, 2005.

[7] NHS Information Authority Homepage, www.nhsia.nhs.uk/def/home.asp, *NHSIA – National Health System Information Authority*, NHS, Great Britain, 2004.

[8] D. Ropeik, and P. Slovic, "Risk Communication: A Neglected Tool in Protecting Public Health", *Harvard Center for Risk Analysis*, June 2003.

[9] UCDAVIS Environmental Health & Safety – Animal Use & Care > Hazard Analysis Tool, ehs.ucdavis.edu/animal/risk/, *UCDAVIS Environmental Health & Safety*, 2004.

[10] WHO Statistical Information System, www3.who.int/whosis/menu.cfm, *World Health Organization Statistical Information System*, World Health Organization, 2004.