A Large Scale Data Mining Approach to Antibiotic Resistance Surveillance

Eugenia G. Giannopoulou[1,2], Vasileios P. Kemerlis[2],
Michalis Polemis[3], Joseph Papaparaskevas[4], Alkiviadis C. Vatopoulos[3],
and Michalis Vazirgiannis[2]

[1]*Department of Computer Science and Technology, University of Peloponnese*
[2]*Department of Informatics, Athens University of Economics and Business*
[3]*Department of Microbiology, National School of Public Health, Athens*
[4]*Department of Microbiology, Medical School, University of Athens*
egian@uop.gr, vpk@cs.aueb.gr,
resistgreece@nsph.gr, ipapapar@med.uoa.gr, avatopou@nsph.gr,
mvazirg@aueb.gr

Abstract

One of the most considerable functions in a hospital's infection control program is the surveillance of antibiotic resistance. Several traditional methods used to measure it do not provide adequate and promising results for further analysis. Data mining techniques, such as the association rules, have been used in the past and successfully led to discovering interesting patterns in public health data. In this work, we present the architecture of a novel framework which integrates data from multiple hospitals, discovers association rules, stores them in a data warehouse for future analysis and provides anytime accessibility through an intuitive web interface. We implemented the proposed architecture as a web application and evaluated it using data from the WHONET software installed in many Greek hospitals that belong to "the Greek System for Surveillance of Antimicrobial Resistance" network. The contribution of the proposed framework is considered to be a standardized workflow aiming at the integration of data produced by various hospitals into a consistent data warehouse and the use of a mechanism that detects hidden and previously unknown patterns on large datasets, in terms of association rules, which can provide surveillance warnings.

1. Introduction

Undoubtedly, antibiotic resistance is a global problem whose origins can be traced to hospitals; places where microorganisms mainly originate and propagate [1]. Early detection of outbreaks can be achieved by applying proactive surveillance [2][3]. Traditional methods of antibiotic resistance surveillance rely on the reporting of resistance rates of various antibiotic/bacterial combinations, incidence rates and annual or semi-annual summary statistics (i.e., antibiograms). Lately, hospital networks have been established for the continuous monitoring of antibiotic resistance. The “Greek System for the Surveillance of Antimicrobial Resistance” (GSSAR) is such a network comprised by more than 40 national hospitals [4]. The data produced by GSSAR are significant in multiple tasks; their analysis facilitates in understanding the trends of resistance, the detection of epidemics, the development of a national antibiotic policy, as well as the further studying of the genetic and molecular mechanisms responsible for the resistance’s emergence [5].

Traditional monitoring methods, such as the antibiograms, mainly lack qualitative results [6]. Moreover, when the surveillance is performed in hospital networks, the issue of real time analyses on the aggregate data, an important function of any surveillance network [7], rises naturally. Thus, the need for a new generation of tools and methods for intelligent data

analysis is inevitable. Recently data mining techniques, such as the association rules, have been incorporated to discover useful patterns of antibiotic resistance from the analysis of hospital laboratory data [1][6][8].

In this work, we present a web based framework that exploits current networking advances to collect the surveillance data from multiple hospitals, perform data cleaning and consolidation, integrate them in a central data warehouse, extract association rules in real time, store the derived results in a central repository and finally provide an intuitive web interface for rules retrieval under multiple criteria conditions. The association rules retrieval mechanism takes into account both the interestingness measures (i.e., support, confidence, leverage) and the attributes in the antecedent or consequent part of the rules (see section 2.1 - Association rules). We implemented our framework as a web-application in order to provide a widely accessible platform for further analysis of large data sets. In our experimental testbed, we used data from the WHONET software [9] (developed by the World Health Organization - WHO) which is functioning in the GSSAR network and is used to store the results of antimicrobial sensitivity tests performed in the hospitals' laboratories.

The remainder of this paper is organized as follows. The necessary background information is provided in section 2. The functional architecture of our framework is presented in section 3 followed by implementation details in section 4. We summarize our work presenting results, discussion and future research directions in section 5.

2. Background

In this section we review some basic definitions related to the association rules' data mining technique and we analyze the principal procedure regarding rules extraction.

2.1. Association rules

The following is a mathematical formulation of the association rules borrowed from [10]. $I = \{i_1, i_2, ...i_m\}$ is a set of items (e.g., hospital, specimen, organism, etc). $D = \{T_1, T_2, ..., T_k\}$ is a set of transactions, (e.g., microbiological exams), where each transaction T is a set of items such that $T \subseteq I$. If transaction T contains X, this is denoted by $X \subseteq T$. An association rule is an "if-then" rule stating that X associates with Y, denoted by $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \varnothing$. The left hand side, LHS, of a rule is called the antecedent part, and the right hand side, RHS, the consequent part. We define *support* of rule $X \Rightarrow Y$ as the probability of a transaction in D to contain both X and Y, (i.e., $X \cup Y$). The *confidence* of a rule $X \Rightarrow Y$ is defined as $support(X \cup Y) / support(X)$. Hence, support is the percentage of transactions containing both parts of the rule, while confidence is the number of cases in which the rule is correct relative to the number of cases in which it is applicable. An interesting rule must at least have support and confidence values greater than the user-specified minimum thresholds. *Leverage* is another index measuring the percentage of the additional cases covered by both the LHS and RHS, above those expected if LHS and RHS were independent. Thus, leverage is defined as $P(RHS \cup LHS) - P(LHS) * P(RHS)$ and represents the unexpectedness of the rule[1]. Finally, *lift* is a measure of the association's importance that is independent of coverage and is defined as $confidence / P(RHS)$.

[1] Leverage ranges between [-1,1]. Values below 0 identify LHS and RHS as independent, whereas values near 1 stand for an interesting association rule.

2.2. Apriori algorithm

Apriori is the seminal algorithm used for association rules mining in a data set [10]. *Itemsets* (i.e., sets of values in the same tuple) with at least minimum support are called *large* in contrast with the rest which are defined as *small*. The algorithm works in two steps in order to discover large itemsets by means of multiple passes over the data. In the first step, it counts the support of individual itemsets and determines which of them are large. In the second step, association rules are generated from the large itemsets found before. The first step is the one considered "cpu-intensive" and accounts for the greater part of the processing time. In order to make the mining process efficient the algorithm exploits the *apriori* principle, the observation that any superset of a small itemset cannot be a *large* itemset. For the sake of completeness, the algorithm is presented in Figure 1[2].

3. System Model

Our work aims at defining a framework that provides data integration services in public health data coming from multiple hospitals. It exploits the networking advantages offered by the World Wide Web, granting real-time information to the users, derived from data mining analysis. In this section, we describe the four main components of the proposed framework, as illustrated in Figure 2. These are: (1) **Data cleaning and transformation**, (2) **Medical data warehouse**, (3) **Association rules extraction engine and warehouse** and (4) **User Interface**.

3.1. Data cleaning and transformation

In the first component of our framework, the public health data are uploaded asynchronously (i.e., at different times) from the hospitals and stored in a temporary database immediately after a fully automated cleaning and transformation procedure. During this middle step, the data are cleaned, grouped and transformed according to well defined templates and rules (e.g., it may be necessary to keep the minimum information describing the hospital wards, group the clinical specimens and microorganisms and transform the date attributes in three-month periods). This way, we ensure the system's flexibility, as more hospitals can join/use the framework without changing their antibiotic storage standards that may be a matter of surveillance policy.

3.2. Medical data warehouse

The medical data warehouse stores and keeps the aggregate information derived from the previous module, in order to provide a unified data view to the mining algorithms and make the association rules extraction an automated/scheduled procedure. Thus, the data uploaded from different hospitals, which are temporarily handled from the previous component (section 3.1 - Data cleaning and transformation), are moved to the medical data warehouse, using a strict schema for the aggregate information. Data integrity is ensured (i.e., safe and no concurrent transfers among the databases) and knowledge extraction is assisted by (1) simplifying the data format on which association rules are to be applied and (2) by providing larger data sets that may reveal new, unexpected relations.

---

[2] Apriori-gen function takes as argument the set of all large (k-1)-itemsets (L_{k-1},) and returns a superset that contains all large k-itemsets [10].

3.3. Association rules extraction engine and warehouse

This component is responsible for revealing and storing non-obvious relations and hidden patterns in public health data by applying the association rules algorithm. Certain parameters that affect the rules' extraction are passed to the knowledge extraction engine (e.g., the hospital id, the time period) and the derived rules are stored in an appropriately designed warehouse, enabling rules retrieval based on multiple search criteria.

1) L_1 = {Large 1-itemsets};
2) **for** ($k = 2$; $L_{k-1} \neq \emptyset$; k++) **do begin**
3) C_k = apriori-gen(L_{k-1}); // new candidates
4) **forall** transactions $t \in \mathcal{D}$ **do begin**
5) C_t = subset(C_k, t); // candidates contained in t
6) **forall** candidate $c \in C_t$ **do**
7) c.count++;
8) **end**
9) $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
10)end
11)Answer = $\cup_k L_k$;

Figure 1. Apriori algorithm

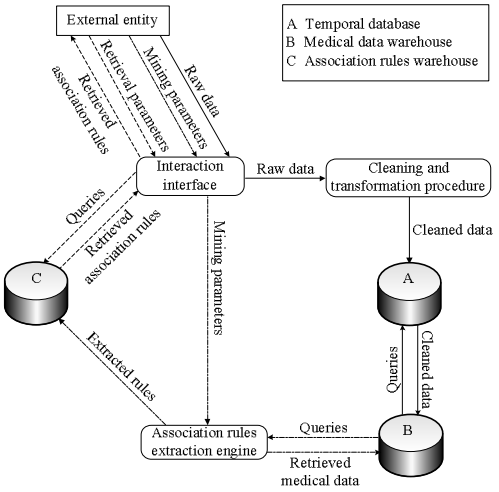

Figure 2. Architectural components and workflow

3.4. User Interface

The user interface is the component responsible for the framework's interoperability. It provides a set of defined control and interaction procedures, among all previous components and aims at minimizing user's errors. The functionalities supported are associated either with the public health data (i.e., uploading, history, preprocessing and statistics) or the association rules (i.e., extraction, history and retrieval). Finally, the web user interface incorporates multiple integrity controls such that overall functionality is efficiently supported.

4. System Implementation and Deployment

We have implemented our framework using open source software tools. The association rules extraction engine was built using Weka's Apriori algorithm [11]. Weka provides a collection of machine learning algorithms, for data mining tasks, developed using Sun's Java platform [12]. We have modified the Apriori implementation to output the mined rules in a certain format that assists further processing. The web application uses dynamic HTML pages [13], produced by the PHP server-side scripting language [14] along with the corresponding database backend (i.e., the warehouses database) based on MySQL [15]. Finally, the application is served to the public using Apache HTTP web server [16].

The data collection that we used comes from the WHONET software installed in several hospitals all over Greece that participate in the GSSAR network. More precisely, our system integrates data from 40 hospitals, 8 different departments (e.g., emergency, medical, surgical, etc.) and 8 different specimen types (i.e., blood, bronchial, genital, pus, skin, stool, urine and wound). The data span a two year period (January 1st, 2003 to December 31st, 2004).

According to domain experts' suggestions, as well as related studies [8], a three-month period is the optimal time interval containing enough tuples to mine for interesting association rules.

Our prototype implementation currently provides partial support to the features described above. In this paragraph, we describe shortly its limitations and the reasons that led to certain design decisions. As far as the preprocessing step is concerned, due to the hospitals' network policy, we only support a certain input data type, based on Microsoft Excel's file format. Thus, the implemented cleaning template is compatible with this type of data only and the transformed data are transferred to the medical data warehouse after the preprocessing step. Association rules are extracted, using the Apriori algorithm, in three stages. In the first stage we extract candidate rules using minimum support and lift thresholds (sample values: 10% and 1.00 respectively). This way, many interesting low-support rules will be included in the results [8]. During the next phase, the produced rules are filtered using minimum leverage and confidence thresholds (sample values: 0.02 and 20% respectively). In the third stage, we truncate the rules that do not align with the format: $\{specimen\ group, hospital, department, period \Rightarrow organism\ group\}$. All the remained mined rules are then stored in the association rules' warehouse for future analysis and retrieval. Finally, as far as the retrieval process is concerned, it enables the user to find subsets of the rules by defining certain measure values (e.g., rules with confidence > 70%) or certain location (i.e., hospital id), department and microbe (i.e., specimen, organism).

5. Results and Discussion

Antibiotic resistance is a serious public health concern. Although traditional methods for retrieving correlations in bacterial identification and sensitivity tests do provide some results, most of them are insufficient to identify underlying associations. Data mining techniques have been incorporated to reveal hidden relations in public health data. This work was motivated by related studies, who aimed at revealing unforeseen patterns. Notice that these efforts' results are strictly tied to certain environments (i.e., one hospital or department). We strongly believe that applying mining techniques on larger data sets will emerge more valid and significant results. Thus, we have presented a framework that enables large scale collection and mining (in terms of association rules) of public health data, at a national level, based on the GSSAR network. New hospitals can easily join this framework and contribute to the derived knowledge. As expected, the majority of the mined rules revealed already known associations. However, interesting results were emerged as well, but due to space limitations we will only refer to a few indicative.

Table 1 summarizes few of our findings. We observe that the rules 2 and 3 show an increasing leverage trend, compared to rule 1. Rule 1 indicates the occurrence of a specific microorganism, *Escherichia coli*, regarding to certain specimen, hospital, department and period. In rule 2, we consider the same microorganism, for the same specimen, hospital and period, but for *all* hospital's respective departments. Although the confidence score of this rule is decreased (this is expected due to the larger data set) when compared with rule 1, the leverage measure is higher. Thus, the confidence that the RHS and the LHS parts of the rule are correlated is stronger (see section 2.1 - Association rules). This observation is generalized by comparing rules 1 and 2 with rule 3. We can see a higher leverage value indicating an even stronger correlation between the LHS and RHS parts when all hospitals and all departments are taken into consideration.

Furthermore, in public health and more specifically in surveillance of antibiotic resistance it is important to discover new associations and patterns before they become widely spread in a hospital or a region. Leverage is a very credible indicator of the public health's importance

of an association rule, since it focuses on the significance of an association and not on its magnitude. For example, rule 4 indicates that in hospital GR032, *Pseudomonas aeruginosa* was isolated more often than expected in ICU (Intensive Care Unit). During period 10-12/2003 has being responsible for 13 additional cases (leverage=0.11, leverage cases=13). Moreover, rule 5 indicates that in hospital GR057, *Staphylococcus aureus* was isolated more often than expected in wounds during period 7-9/2004, that has being responsible for 17 additional cases (leverage=0.1, leverage cases=17).

Table 1. Sample output

	LHS				***RHS***	***Measures***		
#	***Specimen***	***Hospital***	***Department***	***Period***	***Organism***	***Lev***	***Lev cases***	***Conf(%)***
1	Urine	GR004	Med	10-12/2004	E.coli	0.05	24	75
2	Urine	GR004	All	10-12/2004	E.coli	0.1	49	72
3	Urine	All	All	10-12/2004	E.coli	0.11	2264	54
4	All	GR032	ICU	10-12/2003	P.aeruginosa	0.11	13	72
5	Wound	GR057	All	7-9/2004	S.aureus	0.1	17	85

To conclude, our system's evaluation led to conclusions that are tightly connected to the aggregate data collection. Future work will be devoted in using larger data set collections, spanning longer time periods as well as adopting other data mining algorithms (e.g., clustering and predictive modeling) that are expected to contribute to epidemiological studies. Our prototype implementation is available at http://195.251.235.83/en/index.html.

6. References

[1] S. E. Brossette, A. P. Sprague, W. T. Jones, and S. A. Moser, "A data mining system for infection control surveillance", *Methods of information in medicine*, vol. 39, pp. 303-10, 2000.

[2] R. N. Jones, "The current and future impact of antimicrobial resistance among nosocomial bacterial pathogens", *Diagnostic Microbiology and Infectious Disease*, vol. 15, pp. 3-10, 1992.

[3] F. P. Koontz, "A review of traditional resistance surveillance methodologies and infection control", *Diagnostic Microbiology and Infectious Disease*, vol. 15, pp. 43-47, 1992.

[4] WHONET - Greece. http://www.mednet.gr/whonet/

[5] A.C. Vatopoulos, V. Kalapothaki, N.J. Legakis, and the Greek Network for the Surveillance of Antimicrobial Resistance, "An electronic network for the surveillance of antimicrobial resistance in bacterial nosocomial isolates in Greece", *Bulletin of the World Health Organization*, vol. 77, pp. 595-601, 1999.

[6] E. Lamma, M. Manservigi, P. Mello, A. Nanetti, F. Riguzzi, and S. Storari, "The automatic discovery of alarm rules for the validation of microbiological data", *Proceedings of Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP)*, London, UK, September 4th, 2001.

[7] R. P. Gaynes, "Surveillance of nosocomial infections: a fundamental ingredient for quality", *Infectious Control and Hospital Epidemiology*, vol. 18, pp. 475-478, 1997.

[8] L. Ma, F. Tsui, W. R. Hogan, M. M. Wagner, and H. Ma, "A framework for infection control surveillance using association rules", *Proceedings of American Medical Informatics Association (AMIA) Symposium*, Washington, WA, November 9-12, 2003.

[9] WHO - WHONET Software. http://www.who.int/drugresistance/whonetsoftware/en/

[10] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules", *Proceedings of 20th International Conference on Very Large Data Bases (VLDB)*, Santiago de Chile, Chile, September 12-15, 1994.

[11] Ian H. Witten and Eibe Frank, "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco, CA, 2005.

[12] Java 2 Standard Edition Platform. http://java.sun.com/javase/

[13] W3C HTML. http://www.w3.org/MarkUp/

[14] PHP Hypertex Preprocessor. http://www.php.net

[15] MySQL database server. http://www.mysql.com/

[16] Apache HTTP Web Server. http://httpd.apache.org/