Dr Warehouse - An Intelligent Software System for Epidemiological Monitoring, Prediction, and Research

Vladimir Ivančević, Marko Knežević,
Miloš Simić, Ivan Luković
University of Novi Sad, Faculty of Technical Sciences
Novi Sad, Serbia
e-mail: dragoman@uns.ac.rs,
marko.knezevic@uns.ac.rs,
milossimicsimo@gmail.com, ivan@uns.ac.rs

Danica Mandić
University of Novi Sad, Medical Faculty
Novi Sad, Serbia
e-mail: mandiceva88@yahoo.com

***Abstract*—We present Dr Warehouse, an extensible intelligent web-based system for epidemiological analyses. It features a data warehouse containing medical data about registered disease cases and relevant demographical data. There is also a segment of the system that is devoted to presentation and analysis of epidemiological data collected in the data warehouse. The main objectives that we set out for Dr Warehouse include intuitive visualization of epidemiological data, discovery of epidemiological information, and prediction of epidemic dynamics. In the context of epidemiological knowledge discovery, we present a rationale for developing such a system, system architecture of Dr Warehouse, its functionalities, short review of similar systems, and ideas for future development. Furthermore, we describe in more detail choices regarding data modelling, as well as some of the featured predictions and data mining based analyses.**

Keywords-data warehouse; data mining; business intelligence; epidemiological analysis; absenteeism; disease outbreak prediction.

I. INTRODUCTION

Frequent epidemics and various diseases continue to persist in modern world despite great medical discoveries and numerous countermeasures. However, the increase of medical knowledge has helped in the improvement of the overall quality and length of human life. One of the methods for battling diseases includes collection of epidemiological knowledge and its use in the prevention of outbreaks. Our main goal is to contribute to public health by building a software system that could help in the prevention and control of epidemics. This would be possible through the application of results of data analyses featured in the system. Such analyses would be executed on disease case records gathered in the system from various sources.

By following this idea and applying the latest advancements in information technology to epidemiological domain, we created Dr Warehouse – a closed source software system that supports storing of epidemiological data and offers descriptive, as well as predictive, analyses of disease outbreaks. All necessary data are stored in a specially designed data warehouse, while supported analyses include various data visualization techniques, statistical methods, data mining algorithms, and epidemic models. Results of the analyses may be accessed through a rich web client, which offers all of the analyses included in the system, or a mobile device client, which offers a subset of analyses that might be of interest to non-experts. Given the rapid rate of discovery of new analysis methods and epidemic models, we made the system extensible and ensured that new types of analyses may be easily added.

The rest of this paper is organized as follows. Section II looks into our motivation for building such a system. The overview of the system and its components is given in Section III. Some of the predictive analyses supported by the system are presented together with sample results in Section IV. Section V offers a review of similar software systems and their comparison to Dr Warehouse. Section VI includes concluding remarks and ideas for further research.

II. MOTIVATION

A system that could provide its users with a piece of information important in the prediction of epidemics or understanding of disease dynamics would offer many indirect benefits including saving of lives, reduction in treatment costs, and decrease of everyday stress. However, we are also motivated by two more specific reasons: modernization of the healthcare system in Serbia and impact of absenteeism on the economy.

As outlined in the national development strategy [1], the Serbian healthcare system is undergoing a significant transformation. Many segments of that system are being modernized and redesigned to rely more on electronic records as opposed to traditional paper records. Moreover, the expected interconnection of healthcare centres would allow a better electronic access to medical data and consequently better conditions for data analyses, as in the case of the health information system (HIS) for the Serbian Ministry of Defence [2]. In such circumstances, Dr Warehouse could be integrated into the main healthcare system and used for epidemiological analyses. The main system would only be utilized as a data source in the extraction of necessary data, which, after several processing steps, would be stored in the data warehouse within the Dr Warehouse system. Dr Warehouse has been developed also as a potential pilot solution that should demonstrate advantages of using a business intelligence (BI) system in the healthcare domain. It is primarily applicable in activities of institutions that concern themselves with disease prevention in a population, such as institutes of public health.

Besides the fluctuation of labour, absenteeism, which is defined as “failing to report for scheduled work” [3], is the most important parameter that should be monitored by human resources managers in order to increase production potential. This is the case because high absenteeism has negative impact not only on colleagues and superiors, who must cope with greater workloads, but on the profit of a company as well. According to the research from 2009 led by the Chartered Institute of Personnel and Development (CIPD) from Great Britain [4], the most important reasons for the short-term absence from work (4 weeks as maximum) are: colds, influenza, stomach problems, headaches, migraines, injuries of the muscular and skeletal system, as well as pain in the lower back part. Most of these conditions are preventable non-communicable diseases (NCDs) whose rate reduction includes scientifically based cost-effective measures. According to data from population surveys, NCDs are a major health problem in Serbia. Although they are to a great extent preventable, there is no adequate prevention and control of NCDs in Serbia [5, p. 41].

There are several groups of potential users who might benefit from the system that we describe in this paper. Users in healthcare institutions that are dealing with epidemiological data could utilize our software system, which is specially tailored to the epidemiological domain, instead of relying on solutions that are intended for generic statistical analyses. An expected advantage of having a domain-specific system would be an increase in user productivity. Large amounts of data that are typical of modern HISs may be well utilized owing to the well-tried approach incorporated into our system – a data warehouse for data storing and data mining for efficient analyses. In this manner, the main system load may be reduced by running analyses primarily on data stored in the Dr Warehouse system. The second group of users includes scientists whose research is related to epidemiology. By utilizing the Dr Warehouse system, they may create, test, and improve epidemic models through adding, running, and modifying new extensions. New visualization techniques for epidemiological data may be similarly employed and evaluated. Furthermore, the system may also target users who are not medical experts but are interested in latest disease trends, forecasts, or results of some specific analysis.

Bearing in mind the facts concerning the adverse health of the population in Serbia and the “white space” in terms of medical services aimed at predicting occurrence of certain diseases, our decision to develop a system that would allow the use of BI technologies in such a context is both socially and economically justified.

III. System Overview

In this section, we present the system and give an overview of its architecture and functionalities. The featured data warehouse, which represents a foundation for data analyses, is explained in more detail. We also elaborate on the built-in support for adding new functionalities.

A. *System Architecture*

There are four principal components in the system: (i) database server, which has a built-in support for extensibility and includes subcomponents: relational database management system, services for data mining and multidimensional analysis, and services for extracting, cleansing, transforming, and loading data from various sources; (ii) application server, which supports extensibility and acts as an intermediary between database server and clients; (iii) web client application, which supports extensibility and smart card reading; and (iv) mobile device client.

The system may fit into existing HISs and provide various services to other similar solutions. This architecture allows the possibility of having the database server and application server reside at different physical locations. Furthermore, in order to increase the scalability and performance of the system, the data mining and analysis services (currently implemented using Microsoft SQL Server Analysis Services [6]) may be located separately from the database server. In future versions of the system, the architecture may be extended to include terminals that would be publicly available and offer a set of functionalities similar to those in the existing web client application (currently implemented in Microsoft Silverlight [7]).

B. *Data Warehouse*

The data warehouse is modelled using a star schema, which consists of eight dimensions, two of which are role-playing dimensions, and one fact table (Fig. 1). The fact table keeps track of events which lead to absenteeism, disease occurrences and time measured in days that person spent away from duty or workplace. Each dimension represents the context of disease occurrence and absence. Therefore, we can observe these events in the context of time (when an event occurred or ended), gender of the person involved, place where it happened, person’s profession, data source, absence cause, person’s age, and diagnosis that was established. Dimensions concerning diagnosis, place, and time have several hierarchical levels modelled as a fully denormalized structure, which enables multi-level classification of factual data. In the time dimension, we have two hierarchies: one defined as calendar year, quarter, month, and day, and the other one as calendar year, week, and day. The diagnosis dimension has three levels of hierarchy for diagnosis, disease subcategory, and disease category, while community (place) dimension has four levels of hierarchy for community, state, region, and continent. Although the normalization of our schema would remove redundant data and hence become easier to maintain and change, our initial considerations of the schema type led us to choose the star schema. Denormalization, which is typical for the star schema, helped us to reduce the number of foreign keys and to reduce the query execution time. As the system was designed to be used by a wide variety of users, ease of use was one of our priorities. For end users, the star schema is more comprehensible than snowflake schema and less complex queries are needed to satisfy their information needs. Since this is a pilot project, advanced cost-benefit analysis of normalizing our star schema into the showflake schema is a matter of our future work. The unavailability of a larger and more complex absenteeism data set was a major reason for simplifying the initial schema design and focusing on the aforementioned fact and dimensions.

The data warehouse was implemented using Microsoft SQL Server 2008 [8]. It includes the following dimensions: *DimCause*, *DimDiagnosis*, *DimGender*, *DimProfession*, *DimCommunity*, *DimDataSource*, *DimTime*, and *DimAge*. *DiseasePresence* is the only fact table in the system. Each of these tables contains a surrogate primary key which allows us to deal with changes in natural key in a more convenient way and track slowly changing dimensions.

Taking into consideration that the data in the system are expected to reflect the actual state of the health of a population, it is necessary to support acquisition and integration of medical data from multiple sources. We developed a solution within Microsoft Integration Services [9], which allows us to extract, clean, transform, and load (ECTL) the necessary data. We perform incremental extraction, i.e., we consider only data that were added to the HIS of a public health institute after the previous extraction. Extracted data serve as an input for a series of transformations in which we detect and eliminate errors and inconsistencies: (i) different domains of semantically equivalent attributes (as in the case of the attribute *GenderName*); (ii) different encodings of textual data (*ProfessionTitle*); (iii) different granularity of semantically equivalent attributes (*DiagnosisCode*). Diagnosis codes that are used in the source HIS are shorter versions of the codes that are featured in the 10th revision of International Classification of Diseases (ICD 10) [10]. We created a transformation that relies on regular expressions to resolve this issue. In this manner, we extended disease information with the disease name, subcategory and category. At the moment, there is only support for data insertion. Since the data set in the current version of the system is only a sample taken from a HIS, we decided to keep all data in the data warehouse, while leaving the implementation of a deletion policy for obsolete data, data that have little or no impact on the system output, to be included in the future version.

In order to meet the needs for efficient and flexible consumption of valuable information produced by the system, we developed an online analytical processing (OLAP) database, which contains rich metadata. The OLAP cube makes our data organized in a way that facilitates non-predetermined queries for aggregated information. As we used Kimball Method [11] to implement the dimensional model in the relational database, the OLAP design step was a straightforward translation from the existing design. The relational database serves as the permanent storage of the cleaned and conformed data, and feeds data to the OLAP database. Data mining structures and models are stored in the third database, which, together with the OLAP database, resides at the Analysis Server – the primary query server.

C. System Functionalities

The communication between the application server and clients is done via web services. At present, the client applications possess functionalities concerning: access to medical records stored in the data warehouse; access to the data cube and use of some of the cube's advanced analytical operations; execution of advanced analyses and forecasts, as well as result retrieval; services specially tailored for mobile device client; upload of extensions; and their invocation.

The system, whose public resources are available at [12], may be accessed via a web application in which each page groups a number of similar functionalities. Within *Home page*, users may access information about the most common causes of absenteeism for the current month. *Analysis page* contains functionalities regarding the execution of advanced analysis and forecasts that identify the most probable diseases (or causes of work absence) and the most frequent

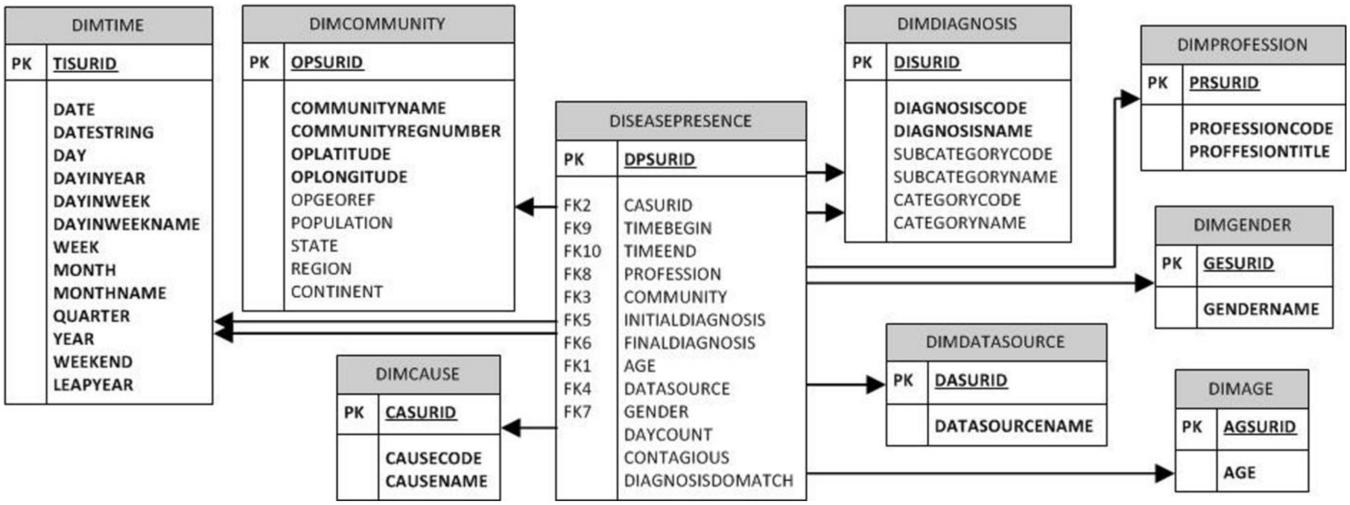


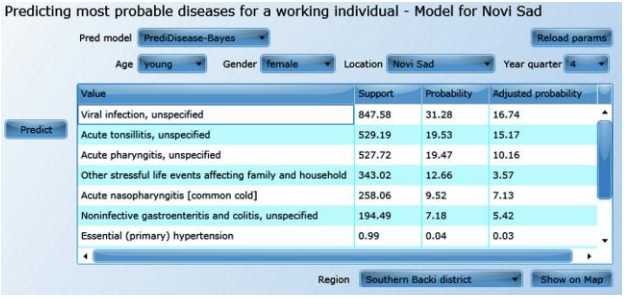
Figure 1. The star schema of the data warehouse.

diagnosis mismatches. Through this page, a user is able to generate predictions concerning a selected subpopulation for a particular quarter of a year. The subpopulation may be specified by selecting an age group, gender, and municipality (Fig. 2). *What about me? page* is a location from which we may generate and retrieve results of the personalized predictions concerning the most probable diseases (or causes of work absence). In order to generate these predictions, all a user needs to do is insert his or her identity card (ID card) into the attached smart card reader and a report is automatically generated. Execution of analytical operations and access to historical data is provided within *Health Reports page*. Users may perform operations such as dice and slice in order to analytically process the available data. *Upload page* offers functionalities regarding uploading of server and client extensions. Within *Extensions page* users may activate and run uploaded extensions.

Some of the aforementioned functionalities are also available via a mobile application for Microsoft Windows Phone [13]. These include disease predictions for a selected location (or the current location of a mobile device) and personalized predictions similar to those featured in *What about me? page* in the web client.

D. *Extensibility*

New functionalities may be added to the system in the form of extensions. The support for extensibility was implemented using Managed Extensibility Framework (MEF) [14]. A user may upload an extension, which then becomes immediately available for use without a need to restart the system. There are two types of extensions: (web) client extensions and (application) server extensions. Both may be uploaded to the application server through the web client. A web client extension is automatically downloaded from the application server to a web client machine, where it is then executed. This is done upon the first invocation of the extension at the client side. Such extension is actually a Silverlight web page that is generally expected to act as a user interface to the built-in or user-added (via server extensions) queries and analyses. On the other hand, server extensions reside on the application server, where they are also executed upon the invocation initiated at the client side. These extensions are functions generally responsible for data operations, analyses, and epidemic models.



Predicting most probable diseases for a working individual - Model for Novi Sad

Pred model: PrediDisease-Bayes

Age: young Gender: female Location: Novi Sad Year quarter: 4

Value	Support	Probability	Adjusted probability
Viral infection, unspecified	847.58	31.28	16.74
Acute tonsillitis, unspecified	529.19	19.53	15.17
Acute pharyngitis, unspecified	527.72	19.47	10.16
Other stressful life events affecting family and household	343.02	12.66	3.57
Acute nasopharyngitis [common cold]	258.06	9.52	7.13
Noninfective gastroenteritis and colitis, unspecified	194.49	7.18	5.42
Essential (primary) hypertension	0.99	0.04	0.03

Region: Southern Backi district

Figure 2. Section from the Analysis page in the web client.

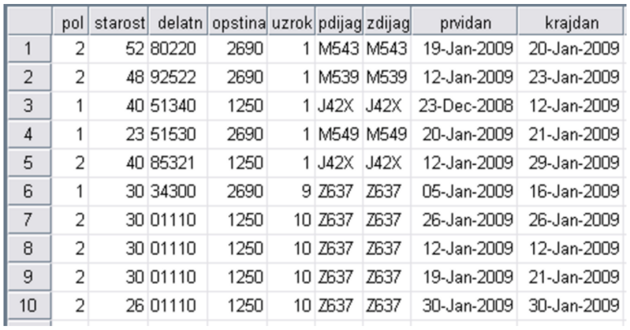
IV. FEATURED EPIDEMIOLOGICAL FORECASTS

In this section, we present two types of epidemiological forecasts that are available in Dr Warehouse: forecasts that rely on data mining and forecasts that rely on compartmental models. In addition to describing a data set that was used, we offer exemplary results of these forecasts.

A. *Data*

Data set used in the testing of the system during the development is acquired from the HIS of The Institute of Public Health of Vojvodina in Novi Sad, Serbia. The obtained sample (an excerpt is featured in Fig. 3) has approximately 8,500 records about workplace absences that ended in 2009. It contains depersonalized information including: gender (represented by the variable *pol*), age (*starost*), municipality code (*opstina*), absence cause (*uzrok*), start (*prvidan*) and end date (*krajdan*) of absence, disease codes for initial (*pdijag*) and final (*zdijag*) diagnosis, and business activity code (*delatn*) of a person involved.

Gender is represented by numbers 1 and 2 referring to the male or female respectively. Business activity code is represented by a five-digit code indicating sector, division, branch and group of a business activity in accordance with the classification of activities as defined by the corresponding law of the Republic of Serbia. Municipality code is a unique identifier of the municipality in which an absence was recorded. The cause of the absence is denoted by numbers from 1 to 12 that respectively correspond to: disease, isolation, accompanying sick person, maintenance of pregnancy, tissue and organ donor, injury at workplace, injury outside of workplace, occupational disease, nursing a child under 3 years, nursing a child over 3 years, care of other sick person, and maternity leave. Initial and final diagnosis codes are obtained by reducing the appropriate diagnosis codes defined by the 10th revision of International Classification of Diseases (ICD 10) to four characters. Codebooks of diseases, business activities, causes and municipalities may be gathered from official Internet sites of organizations that are responsible for their maintenance and distribution. Credibility of the data depends largely on the credibility of data sources. Therefore, we rely on sources that can guarantee the integrity and validity of provided data.



	pol	starost	delatn	opstina	uzrok	pdijag	zdijag	prvidan	krajdan
1	2	52	80220	2690	1	M543	M543	19-Jan-2009	20-Jan-2009
2	2	48	92522	2690	1	M539	M539	12-Jan-2009	23-Jan-2009
3	1	40	51340	1250	1	J42X	J42X	23-Dec-2008	12-Jan-2009
4	1	23	51530	2690	1	M549	M549	20-Jan-2009	21-Jan-2009
5	2	40	85321	1250	1	J42X	J42X	12-Jan-2009	29-Jan-2009
6	1	30	34300	2690	9	Z637	Z637	05-Jan-2009	16-Jan-2009
7	2	30	01110	1250	10	Z637	Z637	26-Jan-2009	26-Jan-2009
8	2	30	01110	1250	10	Z637	Z637	12-Jan-2009	12-Jan-2009
9	2	30	01110	1250	10	Z637	Z637	19-Jan-2009	21-Jan-2009
10	2	26	01110	1250	10	Z637	Z637	30-Jan-2009	30-Jan-2009

Figure 3. Excerpt from a data set used in the generation of predictions.

B. *Forecasts based on Data Mining*

In Dr Warehouse, we utilize three classification algorithms that are supported by Microsoft SQL Server 2008 R2 Analysis Services: decision trees, naive Bayes, and neural network classifier. These classifiers are trained to estimate the individual share of each disease in all work absences attributed to the 15 most common diseases, as determined by examining the available data set, for a selected year quarter and subpopulation, as defined by age group and gender, in a selected municipality. In this manner, we may form coarse predictions of the distribution of the most common diseases in a selected subpopulation. In Fig. 4, we give a set of predictions for male employees in the city of Novi Sad who are between 40 and 61 years old. This example demonstrates how a share of some common diseases in that subpopulation may change throughout a year. These estimates are generated using the naive Bayes classification algorithm for Novi Sad.

Predicted shares indicate that essential hypertension, dorsalgia (thoracic region), and lumbago with sciatica may be causes of a larger percentage of absence in quarters 2 and 3 (spring and summer), while their share substantially decreases during quarters 1 and 4 (winter and autumn). On the other hand, viral infection is most responsible for absences in quarter 4 (autumn).

C. *Forecasts based on Compartmental Models*

Compartmental models are a group of epidemic models that are used to predict dynamics of an epidemic by dividing an analysed population into several compartments (subpopulations) and calculating the changes in compartment sizes given some initial conditions [15-17]. These conditions include sizes of compartments (generally expressed as percentages of a whole population) at a single moment in time. Population compartments correspond to infected, recovered, or some other group of individuals in a population. Furthermore, there are disease-related parameters that are needed in the calculation of changes in compartments sizes: contact rate, recovery rate, death rate, etc. Different models from this family feature different compartments and may be used to obtain forecasts for different diseases. Actual spread of a disease (transition of individuals between different compartments) is modelled by a system of differential equations.

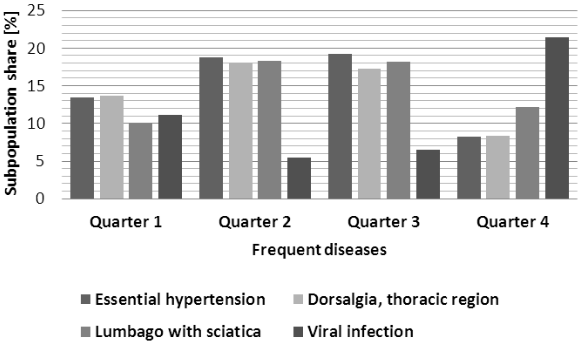


Figure 4. Example of percentage disease shares for male employees in Novi Sad aged between 40 and 61 years, as predicted using a naive Bayes classifier.

As an example of how Dr Warehouse may support standard epidemic models, we implemented the SIR (Susceptible/Infected/Recovered) model as an extension pair consisting of: (i) a client extension, which is used to set parameters, invoke model execution, and present results; and (ii) a server extension, which is an invoked function that numerically solves the system of equations and prepares results for the client side.

The implementation is an adaptation of the model version featured in [18]. The name of the model is derived from the three compartments that are used to model a population struck by a disease: susceptible (S), infected (I), and recovered (R). A susceptible individual from the S compartment may become infected through contact with an infected individual from the I compartment, while an infected individual may become a member of the R compartment after a recovery period. A rate at which a disease is transmitted from an infected to a susceptible individual is the contact rate β, while a rate at which an infected individual recovers is the recovery rate γ. Actual values for the rates β and γ depend on a disease that is being modelled. Three ordinary differential equations describe the dynamics:

$$dS / dt = -\beta I S, \quad (1)$$

$$dI / dt = \beta I S - \gamma I , \quad (2)$$

$$dR / dt = \gamma I. \quad (3)$$

Our implementation approximates the solution of this system by using the 4th order Runge-Kutta method for solving a system of ordinary differential equations. In Fig. 5, we give an example of a prediction that was generated by a chronological simulation using our implementation of the SIR model. The presented chart demonstrates a typical situation when equilibrium in a population is gradually reached after a peak in the number of infected individuals.

Other compartmental models may be implemented in a similar manner. The only differences would be changes in a set of differential equations that model a disease and addition of new parameters or compartments. By adding the support

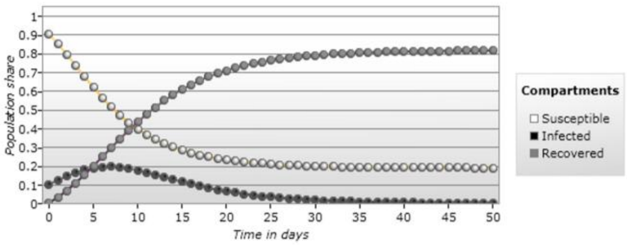


Figure 5. Example of a disease forecast obtained using the SIR model.

Copyright (c) IARIA, 2013. ISBN: 978-1-61208-247-9

for the SIR model in the form of a pair of extensions, we have demonstrated that Dr Warehouse may be used to predict the rate of spread of any disease for which there is an adequate compartmental model.

V. Related Work

There are many software systems for epidemiological analyses and monitoring. One group of such systems provides mostly statistical procedures that are often used in epidemiology. Open Source Epidemiologic Statistics for Public Health (OpenEpi) [19] is an example of a freely available system that may be run in a web browser [20] because it is implemented in HyperText Markup Language (HTML) and JavaScript. It focuses on statistical calculations: calculation of confidence interval and sample size, estimation of power for different types of studies, execution of various statistical tests, etc. Another free solution is WinPepi [21], which is a set of desktop applications that are similar to OpenEpi and offer many statistical procedures that are useful in epidemiology. When compared to Dr Warehouse, both OpenEpi and WinPepi are projects of a narrower scope because they ignore data storage and management. Furthermore, they put emphasis on statistics and a large number of calculation modules whose input is mostly a small set of summarized values. Unlike Dr Warehouse, they do not support data mining, visual representation of data, epidemiological maps, nor user extensions. However, the source code of OpenEpi may be directly modified to include new procedures.

The second group of epidemiological systems includes data storing and manipulation capabilities in addition to analysis procedures. Epi Info [22] is one such example of a desktop software application with a wider range of functionalities than OpenEpi and WinPepi. What sets it apart from other software systems for epidemiology is the support for form creation. A user may design custom forms through an integrated editor and later use them for data entry. Besides basic and advanced statistical procedures, this system supports data import and export, as well as basic data selection and transformation. It has good data visualization capabilities and offers various types of charts, tables, and even map overlay. Its main strengths with respect to Dr Warehouse are support for form creation, direct data entry, data transformation and data import/export for various types of data sources. However, there is a conceptual difference between these two systems regarding data storage. Epi Info is a tool that may be used over any data (in the supported file or database format) and, therefore, provides transformation functions, which a user utilizes in order to prepare data for analyses. On the other hand, Dr Warehouse features a data warehouse with a fixed set of facts and dimensions, and a carefully designed ECTL process, which is automatically executed. Therefore, there is generally no need for manual data import and transformation because data preparation is done automatically. In other words, Dr Warehouse may be seen as a more specialized and more automated solution in which the data warehouse has a prominent role. Our system relies on a strong dependency between the data warehouse schema and analyses, which helps to simplify the analysing process. This, in turn, alleviates much of the burden concerning data preparation, which is usually the longest activity in analysis projects. Some of the main features of Dr Warehouse that Epi Info lacks are data mining procedures and the support for adding user extensions. We consider data mining to be an essential part of the system because, unlike most statistical procedures, it is well suited for analysing large quantities of data that are efficiently stored in a data warehouse. We may summarize this comparison by generally classifying Epi Info as a solution that offers a fixed set of analyses for any set of data attributes and Dr Warehouse as a solution that features a fixed set of data variables but an extensible set of techniques for data presentation and analysis.

The third group consists of typically web-based systems that focus on epidemiological monitoring and publicly presenting latest disease outbreak data for different regions throughout the world. They primarily rely on data from numerous Internet-related sources, which may be informal or official. HealthMap [23] provides a world map with the latest information on outbreaks by automatically collecting and integrating data mostly from several online news sources and reports from eyewitnesses and officials. There is also a mobile version of the system with similar functionalities. Another web system with a support for mobile devices is Outbreak Watch [24]. It does real-time analyses of data in social networks by evaluating keywords that are considered to be indicators of outbreaks. In this manner, the system tracks changes in the number of reports concerning relevant diseases. Google Flu Trends [25] was created as an attempt to estimate actual flu activity in various countries by analysing aggregated Google search queries that are related to flu. Since there is a relationship between an actual number of flu cases and search queries about flu, as confirmed by the overall match between the official surveillance data and the calculated estimates, this service offers near real-time results, which may help in preparing a response to a flu outbreak. Dr Warehouse is similar to these systems, as it may offer latest epidemiological data and forecasts in the form of charts, tables, and maps. In addition to supporting web access, it also features a mobile version with a selected set of services. On the other hand, the principle difference lies in the selection of data sources. The three monitoring systems use data that are available on the Internet (HealthMap and Outbreak Watch) or from web search queries (Google Flu Trends), while Dr Warehouse displays only data present in the data warehouse, which was planned to include credible data collected in healthcare institutions. However, the ECTL process in Dr Warehouse may be extended in the future to include data from public web sources.

When compared to the three aforementioned groups of epidemiological software, Dr Warehouse is a complex system that possesses traits typical of all three because: (i) it may offer any statistical procedure that has been added as an extension; (ii) data management is one of the key segments of the system; and (iii) collected data are constantly available to users via web and mobile client, which makes the system suitable for epidemiological monitoring.

VI. Conclusion and Future Work

In this paper, we introduced a software system that may be used in epidemiology for data collection, data mining, analyses and research. We expect that this system may have an important role in the activities concerned with epidemic control and better understanding of temporal and spatial disease patterns. In addition to a general description of the system's structure and components, a special attention was given to the implementation of its data warehouse and data analyses. The presented examples of analyses illustrate just some of the results that may be obtained through the current version of the system. In order to support application of the latest advances in epidemiology and evaluation of new epidemic models, we incorporated extensibility that allows addition of new functionalities to the system.

There are numerous ideas for future work and research on the presented system. We may modify the existing analyses and make them more generic so that they could support a greater number of queries. The data warehouse schema may be altered and extended in order to support additional analyses. We may also enforce a strict security policy by introducing user roles and separating the set of functionalities into subsets better suited for various user categories. A new version of the system could be implemented using open (and free) technologies, which could lead to a creation of a completely open version of the system. Due to the prominence of spatio-temporal and epidemiological data in Dr Warehouse, best practices from geographic information systems and constraint databases are topics also worth exploring in the future. Furthermore, significant additions to the system would be construction of an epidemiological knowledge base, which could be regularly updated or consulted during data analyses, and creation (or selection) of a convenient ontology. In this manner, the semantics may be expressed and the new version of the system could communicate with other systems that follow the idea of the Semantic Web.

Acknowledgement

The research was supported by Ministry of Education and Science of Republic of Serbia, Grant III-44010. The authors are most grateful to The Institute of Public Health of Vojvodina in Novi Sad for the provided absenteeism data sample and valuable comments.

References

[1] *Strategija razvoja informacionog društva u Republici Srbiji do 2020. godine* [The Strategy for the Development of Information Society in the Republic of Serbia until the Year 2020], (in Serbian), Službeni glasnik Republike Srbije, vol. 51, 2010.

[2] M. Fimić, M. Radulović, I. Vulić, and S. Atanasijević, "Zdravstveni informacioni sistem Ministarstva odbrane Republike Srbije – generičko rešenje za integraciju institucija" [The Health Information System of the Ministry of Defense of the Republic of Serbia – A Generic Solution for Institution Integration], (in Serbian), in Proceedings of YU INFO 2012, pp. 511-516.

[3] G. Johns, "absenteeism," in *The Blackwell Encyclopedia of Sociology*, G. Ritzer, Ed. Oxford, UK: Blackwell Publishing, 2007, pp. 4-7.

[4] "Absence management 2009 – Survey Reports - CIPD," http://www.cipd.co.uk/hr-resources/survey-reports/absence-management-2009.aspx [Dec. 7, 2012].

[5] Đ. Jakovljević and P. Mićović, *Zdravstveno stanje i zdravstvene potrebe stanovništva Srbije* [Health Status and Health Needs of the Population of Serbia], (in Serbian), http://www.palgo.org/files/leaflet/brosura_zdravstvo.pdf [Dec. 7, 2012].

[6] "SQL Server Analysis Services," http://technet.microsoft.com/en-us/sqlserver/cc510300.aspx [Dec. 7, 2012].

[7] "Microsoft Silverlight," http://www.microsoft.com/silverlight/ [Dec. 7, 2012].

[8] "Microsoft SQL Server," http://www.microsoft.com/sqlserver/ [Dec. 7, 2012].

[9] "Microsoft Integration Services," http://msdn.microsoft.com/en-us/library/ms141026%28v=sql.105%29.aspx [Dec. 7, 2012].

[10] "International Classification of Diseases," http://www.cdc.gov/nchs/icd/icd10cm.htm [Dec. 7, 2012].

[11] J. Mundy, W. Thornthwaite, and R. Kimball, *The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset*, 2nd ed., Indianapolis, IN: Wiley Publishing, Inc., 2011.

[12] "Dr Warehouse," http://www.acs.uns.ac.rs/sr/node/237/1140/ [Dec. 7, 2012].

[13] "Microsoft Windows Phone," http://www.microsoft.com/windowsphone/ [Dec. 7, 2012].

[14] "Managed Extensibility Framework," http://msdn.microsoft.com/en-us/library/dd460648.aspx [Dec. 7, 2012].

[15] W.O. Kermack and A. G. McKendrick, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London*, vol. 115, no. 772, pp. 700-721, August 1927.

[16] R.M. Anderson and R.M. May, "Population biology of infectious diseases: Part I," *Nature*, vol. 280, no. 5721, pp. 361–367, August 1979.

[17] R.M. May and R.M. Anderson, "Population biology of infectious diseases: Part II," *Nature*, vol. 280, no. 5722, pp. 455–461, August 1979.

[18] M.J. Keeling and P. Rohani, *Modeling Infectious Diseases in Humans and Animals*. Princeton, NJ: Princeton University Press, 2007.

[19] K.M. Sullivan, A. Dean, and M.M. Soe, "OpenEpi - a web-based epidemiologic and statistical calculator for public health," *Public Health Reports*, vol. 124, no. 3, pp. 471-474, May-June 2009.

[20] "Open Source Epidemiologic Statistics for Public Health," http://www.openepi.com/ [Dec. 7, 2012].

[21] J.H. Abramson, "WINPEPI updated: computer programs for epidemiologists, and their teaching potential," *Epidemiologic Perspectives & Innovations*, vol. 8, no. 1, pp. 1-9, February 2011.

[22] "Epi Info™ - Community Edition," http://epiinfo.codeplex.com/ [Dec. 7, 2012].

[23] "HealthMap," http://www.healthmap.org/ [Dec. 7, 2012].

[24] "Outbreak Watch Social Biosurveillance Network," http://www.outbreakwatch.com/ [Dec. 7, 2012].

[25] "Google Flu Trends," http://www.google.org/flutrends/ [Dec. 7, 2012].