Assisting School Units Management with Data Mining Techniques and GIS Visualization

John Garofalakis, Antonios Maritsas and Flora Oikonomou
Computer Engineering and Informatics Department, Patra's University, University Campus, Patra, Greece

Keywords: Educational Data Mining, Geographic Information Systems, Visualization, Decision Support, Clustering, Classification, Epidemy Spread.

Abstract: Educational Data Mining (EDM) has emerged as an interdisciplinary research area that applies Data Mining (DM) techniques to educational data in order to discover novel and potentially useful information. On the other hand, Geographic Information Systems (GIS) are ones designed to manage spatial data and related attributes and can be used for assisting decision support. This paper proposes an innovative use of DM and visualization GIS techniques for decision support in planning and management of Greek public education focused on high risk groups such as young children. The developed application clusters school units with similar features, such as students' and teachers' absences, and represents them on a map, enabling user to make decisions being aware of geographical information. Afterwards, based on real data stored during epidemic spread periods, such as the H1N1 flu pandemic during 2009, the application predicts whether a school should be opened or closed considering students' and teachers' absences of a specific time interval.

1 INTRODUCTION

Educational knowledge representation has emerged as a burgeoning new area in the research landscape. The need to unambiguously describe knowledge deriving from educational data has given rise to the development of methods for exploring unique types of data which come from educational context.

Increasingly, the use of educational software and state databases of student information has created huge depositories of data. Moreover, the use of Internet in education resulted in the context of e-learning in which large amounts of information are ubiquitously available and boundlessly generated (Castro et al., 2007). Educational Data Mining (EDM) is the application of Data Mining (DM) techniques to educational data, and so, its objective is to analyze these type of data in order to resolve educational research issues (Barnes et al., 2009). The techniques which EDM uses convert raw data deriving from educational systems into usefull information which could possibly have an impact on educational research and practice.

On the other hand, nowadays people's daily habits, from the common ones to the most complex, are most of the times interwoven with the concept of space. Practically, all the decisions made in government, business or scientific level are influenced or determined by geographical characteristics. Decisions are made after evaluating several data which are characterised as information and are space connected (Kapageridis, 2006).

A Geographic Information System (GIS) is an information technology which stores, analyses, and displays both spatial and non-spatial data (Parker, 1988). Or, as Cowen states (Cowen, 1988), a GIS is a decision support system involving the integration of spatially referenced data in a problem-solving environment. Nowadays, there is an emerging interest for these systems which are designed to manage spatial data and related attributes and are also used for decision support.

This paper describes an application of data mining and decision support in the educational field within a project called GeoMapping. The goal of GeoMapping is to assist the knowledge management by enabling user to make decisions being aware of geographical information. The application analyzes data from Greek school units of primary and secondary education and it geographically represents them. It uses data mining techniques in order to cluster school units with similar features, such as students' and teachers' absences, and represents them on a map.

Garofalakis, J., Maritsas, A. and Oikonomou, F.
Assisting School Units Management with Data Mining Techniques and GIS Visualization.
DOI: 10.5220/0006317603310338
In *Proceedings of the 9th International Conference on Computer Supported Education (CSEDU 2017) - Volume 1*, pages 331-338
ISBN: 978-989-758-239-4

Afterwards, based on absences' data stored from school year 2009-2010 due to H1N1 virus, the application predicts, whether a school should be opened or closed.

The paper is organized as follows. Section 2 presents background information concerning specific needs leading to the decision to implement the application. Section 3 presents the application. The results of applying data mining and decision support techniques, and the visualization of the results, are presented in Section 4. Section 5 concludes by summarizing the application's results and by presenting plans for future work.

2 BACKGROUND

Since the late 1990s government agencies around the world have embraced the digital revolution and placed a wide range of materials on the Web including publications, databases and actual online government services (West, 2002). The construction and management of e-government systems is becoming an essential element of modern public administration (Torres et al., 2005) in order to enhance the access to and delivery of government information and service to citizens. Moreover, e-government facilitates the decision making process of national interest matters.

Prospective monitoring of public health is a crucial element of the strategies used for the control of diseases at a national level. The effective collection and analysis of data feels necessity for specialized methods and tools (What is Epidemiology All About, 1999; Burkom et al., 2005) which Information and Communication Technologies (ICT) achieve it effectively.

Back in 2009, the wide fast spread of the H1N1 virus was a specific case where ICTs assisted authorities at their effort to take the needed measures in order to minimize the virus's evolution. Several applications collected data concerning factors directly or indirectly related with H1N1 spread. One of these applications was the Absences System (part of the Survey System) of the Greek Ministry of Education (Garofalakis et al., 2011).

The latter system recorded daily absences from school units in Greece of primary and secondary education. Studies have shown that children are prone to viruses and the monitoring of this age group can provide useful information about a virus's spread in a country. The aforementioned application assisted the Greek Ministries of Education and Health on their effort to epidemiologically monitor H1N1 evolution in Greece and compensated to extract conclusions for taking public health measures in the country.

However, the application neither provided any integration with a web GIS adaptive interface for better visualization of the stored data nor performed any data mining on the data in order to draw helpful conclusions. Public health care is a knowledge intensive domain in which neither data accumulation nor data analysis can be lucrative without using knowledge about both the problem domain and the data analysis process. This indicates the usefulness of integrating decision support techniques with data mining in order to create effective decision support models. Furthermore, the use of visualization techniques facilitates the knowledge management and improves the whole decision support process.

Information visualization uses graphic techniques to help people understand and analyse data (Mazza, 2009). Visual representations and interaction techniques take advantage of human eye's broad bandwidth pathway into the mind to allow users to see, explore, and understand large amounts of information at once (Romero and Ventura, 2010). The objective of data visualization is to highlight useful information and support decision making (Romero and Ventura, 2010).

There are several systems which attempt to visualize educational data. For example, SAMOS system (Juan et al., 2009) monitors students' and groups' activities in collaborative e-learning by using spreadsheets on data contained in server log files. Or GISMO (Mazza and Milani, 2004) generates graphical representations of its data which can be explored by course instructors to examine various aspects of distance students.

Thus, the presented application uses the data stored at the Absences System and by applying descriptive and predictive data mining methods in combination with visualization techniques, achieves to improve the management of data and knowledge. This goal is approached in two phases: first, there is a data analysis with data mining techniques and second the results of data mining are evaluated in order to facilitate decision support. In the second phase, the geographical representation of the data is the one which enriches the decision support process.

3 THE SYSTEM

The GeoMapping system is a web application and it is accessed after authentication through a browser. The authenticated users belong into two levels with different privileges. The first level consists of

administrators, who can access all the stored data and all the application's menus, whereas the second level consists of simple users, who can access the stored data of their area of interest and have limited access to application's menus. The application is personalized based on the user's level. In the following paragraphs, there is an extent description of the structure of the application. In order to better understand the system, there will be a presentation of the Absences system upon which the application has been based for data retrieval. Afterwards, there will be an analysis of the presented system.

3.1 Data Retrieval System

The Absences System (Garofalakis et al., 2011) was part of the Survey System (Garofalakis et al., 2007) which was used to record data about students in school units and teachers employed either in school units or in administration offices. The whole application was integrated in the existing Resources Management System of the Greek Ministry of Education and it operated over the Greek School Network (Greek School Network, 2016) utilizing its access and its security policies. The Absences system started on October 2009, leading to important results assisting in real time the relevant Greek Authorities responsible to deal with the H1N1 spread. The latter system was used until January 2014 when it was replaced by myschool system (myschool, 2016).

Aim of the Absences System was to assist the epidemiology monitoring of H1N1 epidemic evolution in Greece, in the whole country and per region leading to the extraction of important conclusions for taking precautionary measures against the flu (e.g. temporary school closures). The system required the daily record of students' and teachers' absences and the school's state (opened or closed). If the school was reported as closed the user had to declare the reason for its closure (due to flu symptoms or holiday or other reason). Furthermore, if a school unit had a lot of students' absences then the school would not function for a few days in order to prevent the spread of the flu virus.

The system assisted the Greek Ministries of Education and Health on their effort to support the epidemiological monitoring of H1N1 evolution in Greece. Although it provided some statistic results concerning absences data, neither performed any data mining techniques nor geographic representation of the closed school units. However, the data collected from this system constituted the data retrieval system for GeoMapping application in which data mining techniques, geographical representation and prediction techniques were performed in order to assist decision making process.

3.2 The GeoMapping Application

The GeoMapping system constitutes an extension of the Absences system since it retrieves data from the latter and applies several data mining techniques to it for drawing conclusions concerning the epidemiological monitoring of viruses.

As a web application, the system requires user authentication. There are two user levels, administrators and simple users, and the application is personalized based on their level. Administrators can access all stored data and menus whereas simple users have limited access to menus and can access only data related to their area of interest. Moreover, administrators perform account management to all user accounts thus enabling them to change their data, their user level or their status (active/inactive) allowing them to deactivate their account.

Apart from the aforementioned functionalities, the application data mines the absences' data entered by school units, depicts results on a map and can predict a school unit's state (open/closed) based on similar school units' state. The school units we are mainly interested in are those which are thought of as high risk units (i.e. school units with the biggest percent of absences for a specific characteristic). Those high risk units are considered as the best candidates for viruses' spread. The preceding functionalities of the application will be described in the subsequent paragraphs.

3.2.1 Geocoder

The objective of this application was the usage of data mining techniques in an educational context and the representation on a map of the resulted information. The data used in this application was retrieved by the Absences system. Nevertheless, the school unit's addresses in the Absences system did not contain any geographic coordinates and in order to illustrate data mined information on a map, some correlation had to be performed on school units' addresses with geographic coordinates. This correlation was implemented with the usage of a geocoder subsystem which converted the necessary data for all school units stored in the application (Figure 1).

Geocoding is usually synonymous with address matching (Drummond, 1995; Vine et al., 1998; Bonner, et al., 2003) highlighting its prevalent use of transforming postal addresses into geographic representations. In this application geocoding is

performed by using google maps api. The api on the one hand allows the geocoding service usage by using an HTTP request but on the other hand there are some limitations. The api can serve up to 2500 free requests/day with a maximum of 50 requests/sec. Thereby a geocoder subsystem was developed in which a timer delay process maximized api's usage and allowed more requests per second. When the geocoding process is complete, the geocoder returns a JSON array with the geographic coordinates of the postal address which are stored in our system's database. Those geographic representations are then used in order to place markers on a map and depict the conclusions drawn from data mining process.

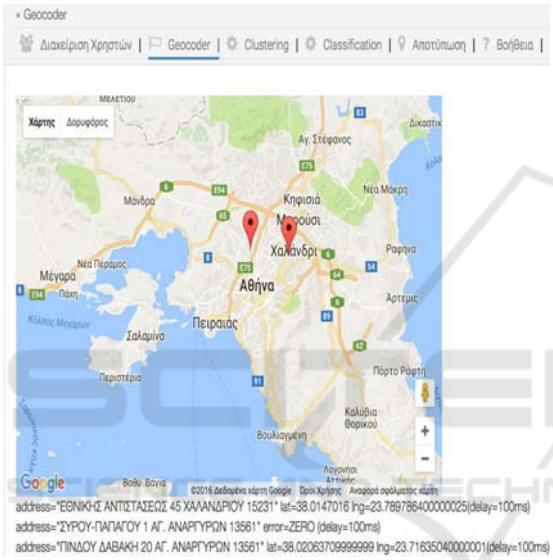

Figure 1: Geocoding operation.

3.2.2 Data Mining

The data mining process from the absences' data is performed by using predictive and descriptive algorithms. The predictive data mining algorithms result in models that can be used for prediction and classification whereas descriptive data mining algorithms result in finding interesting patterns in data, like associations, clusters and subgroups.

In GeoMapping application, we mainly used descriptive data mining methods and combined them with google maps' visualization and decision support techniques to improve the management of data and knowledge at the Absences system and subsequently at the Greek Ministries of Education and Health. In order to achieve the latter, we firstly analysed the available absences' data with data mining techniques and secondly, we used the results of data mining as input for the decision support techniques.

In order to detect similarities between school units with similar absences, two different clustering methods were used: partitional clustering method and density-based clustering. For the first method K-Means algorithm was selected whereas for the second method we used DBSCAN algorithm. Apart from the clustering method, the user in our application had to select the appropriate sample in which data mining would be performed. The samples were the students' age, the students' absences percentage, the teachers' absences percentage and the school unit's absences percentage. At the end, we compared the results provided by those algorithms and we found that the conclusion was the same regardless of the data mining algorithm.

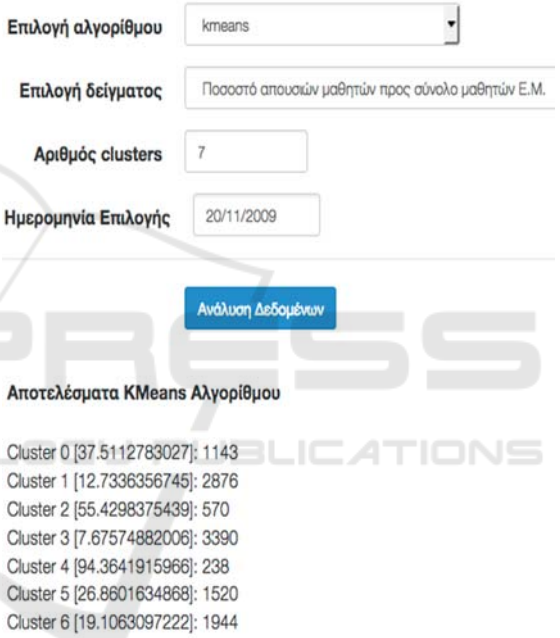

Figure 2: Using K-Means algorithm to cluster school units based on students' absences percent on cold winter day.

An illustration of the clusters, generated by data mining methods is given in Figure 2. In this example, K-Means algorithm is selected in order to cluster the school units during a cold winter day (20/11/2009) when it is expected to have a great number of absences. The sample for the K-Means algorithm is the students' absences percent and the number of the desired clusters is 7. It should be mentioned that after several trials on our dataset we found that for our context the optimal number of clusters is 5 or 7 clusters of data. Apart from the above, we also choose a date in which the clustering will be performed since we have a daily recording of absences. As a result, the K-Means algorithm returns 7 clusters of school units.

Cluster 0, for example, means that 1143 school units were found in which students' absences percent reached 37.5%. In this example, we can see that on 20/11/2009 the majority of school units (3390 units) had a small number of students' absences while a handful of school units' (238) had an absences percent close to 94%. Those school units with the largest percent of absences are the ones we are mainly interested in and this kind of information is stored in our database in order to use it later for visualizing the high risk units. The value of 94% of students' absences constitutes our threshold value which is used in our decision support part of our application.

If the data mining process is performed on a hotter spring day (i.e. 15/04/2010) then the expected number of absences will be significantly smaller compared to those of a cold winter day. In Figure 3 we execute the K-Means algorithm for students' absences on 15/04/2010. In this example, we can see that the majority of school units' (6259 units) had a small number of absences and a minor number of units (8 units) had a students' absences percent close to 98%. This insignificant number of school units will not be taken into account for high risk schools with a possible virus spread.

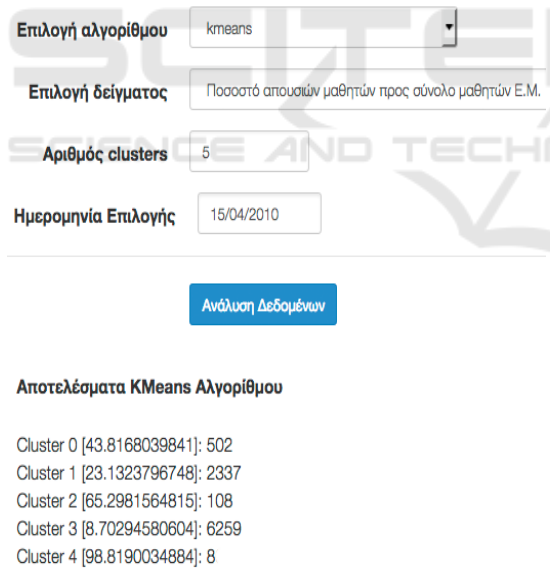

Figure 3: Using K-Means algorithm to cluster school units based on students' absences percent on hot spring day.

3.2.3 Decision Support

By using the results from data mining algorithms, which are basically those school units with the greatest percent of absences for a specific sample, the application creates a selection threshold. This threshold will help to distinguish the high risk units. It has to be mentioned that in order to create the threshold, the application takes into account not only students' but also teachers' absences.

Afterwards, several queries are performed to the application's database in order to discover those school units with absences' percent close to the threshold. The database queries have as a parameter the threshold value for students' and teachers' absences that was calculated in the previous data mining step. With the utilization of the threshold value it was made possible to discover the high risk school units for a specific day from the total school units which are recording absences on a daily basis.

Subsequently, GIS data can be used to visualize those high risk school units in Greece. Instead of raw data visualization, presented by markers of school units, we implemented some processes that allow markers to depict information concerning the high risk school units. That information regards the school unit's name and description.

This leads to the development of a map, which enables the visualization of areas of Greece with high risk school units.

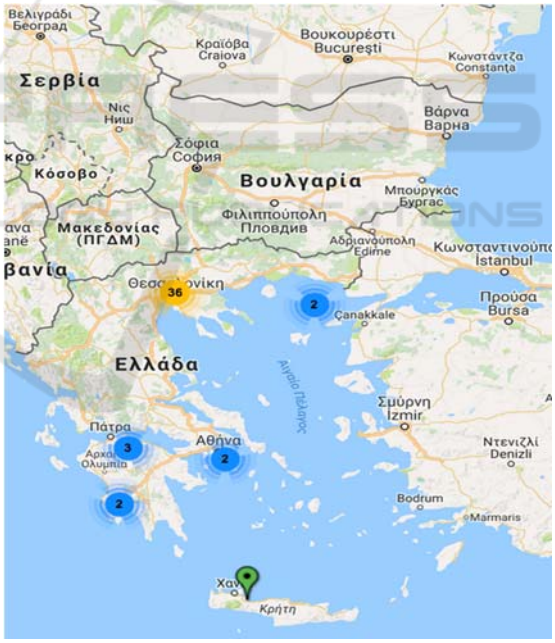

Figure 4 Visualizing high risk school units during a winter day (22/01/2010).

In the example depicted in Figure 4, the decision maker can see that in the area of Thessaloniki a lot of school units had a high number of students' and teachers' absences.

The validation of the acquired results was juxtaposed with the results provided by the "Greek

center of monitoring and prevention of diseases – KEELPNO" (KEELPNO, 2010). KEELPNO provided weekly statistics which depicted the H1N1 virus spread to all regions of Greece. Thus, by examining the graph in Figure 5 which shows the students' absences per region in Greece, it can be seen that GeoMapping application correctly shows a cluster of high risk school units in the area of Thessaloniki during winter and more specifically on 22/01/2010 (North Greece region, brown colored line on Figure 5).

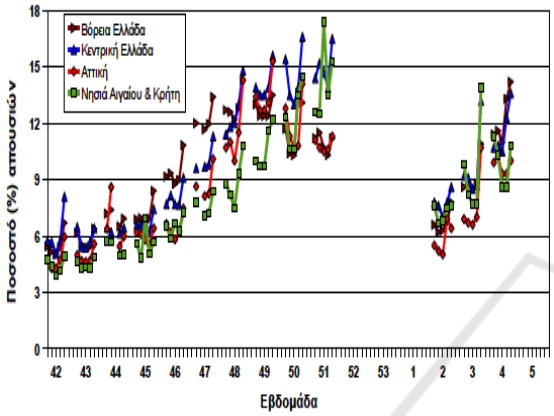

Figure 5: Students' absences per region in Greece (until 4th week of 2010) provided by KEELPNO.

3.2.4 Prediction

Apart from the clustering algorithms which were used in order to discover school units with similar absences results, the application implemented a predictive data mining algorithm for proposing the status of a school unit (whether should remain opened or closed). The classification method which was chosen in order to predict a school unit's state used a training set of absences data retrieved by the Absences system for the first three months of its usage. The algorithm that performed the classification was the ID3 decision tree algorithm.

The ID3 algorithm was selected due to its advantages. It can create a decision tree with the smallest number of steps, it examines the whole database and the choices of one level benefit the other levels.

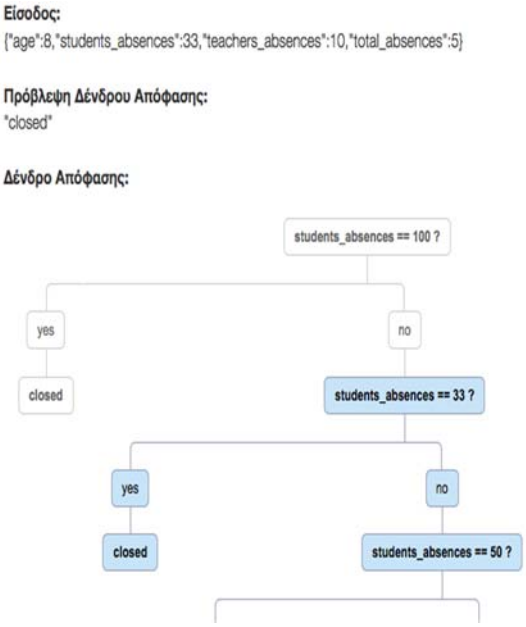

Figure 6: Classification algorithm example.

In our application, the user enters some input data for classification and the system responds whether a school unit should be opened or closed. The input data are the students' age, the students' absences percent, the teachers' absences percent and the total absences percent. For example, the application's user wants to see whether a primary school unit with some students' absences should remain open or closed for precautionary measures. In Figure 6 we can see that the system's response is that the school unit should close for precautionary measures if the age group is 8 years old, the students' absences reach 33%, the teachers' absences reach 10% whereas the whole population absences reach 5%. This kind of information is extremely helpful since it can predict a school unit's state and facilitate the decision making process in order to take precautionary measures (i.e. vaccination to the areas where high students' absences are monitored).

4 FUTURE WORK

The developed application, as any other application, can be further improved driven by results' accuracy and user friendliness. Therefore, future extensions could be the use of other clustering and classification algorithms with more advanced features or the more thorough inspection of the clustering and classification results with evaluation metrics for ensuring the validity and the accuracy of the returned

data. Another improvement could be the maps' enrichment in order to include thematic maps for depicting for example demographic data such as population density.

5 CONCLUSIONS

The use of data mining and decision support techniques, including GIS visualization, can lead to better results in decision making, can improve the effectiveness of developed applications and enables interfering with new types of problems that have not been addressed before.

In the GeoMapping system, a web application was implemented in which data mining techniques are applied to school units of primary and secondary education in Greece and the results of those techniques are geographically represented on a map.

Aim of GeoMapping system is to offer a data mining model which performs two operations. Firstly, the clustering methods group similar school units together based on their students' and teachers' absences. Afterwards, a map representation takes place allowing the user to make a decision taking into account the geographic information. The knowledge of this information is extremely critical in epidemical spread periods when a disease outbreak to a large number of people in a given population within a short period of time necessitates governmental decision making and measures taking. The second operation is related with the prediction of a school unit's state (open/closed) based on stored data. This operation can be used by school units' principals or education executives who are directly in charged with taking decisions for unit's state.

The data used in GeoMapping application was retrieved from the Absences system and is related with all school units in Greece. This characteristic reinforces the completeness and the validity of the returned results.

It has to be mentioned that the results taken from the clustering operation are accurate enough as it can be seen from the comparison made with the official data which was given by KEELPNO for that period of time (school year 2009-2010) (Figure 5). More specifically, the official data showed that many school units of primary education were closed in North Greece region. This comes to an agreement with the results provided by GeoMapping (Figure 4) in which a high risk school units' cluster is depicted in Thessaloniki, North Greece region.

Conclusively, the innovation of this application compared to others is the combination of data mining methods in an educational context and the results' GIS visualization on a map. When critical situations appear, such as a virus outbreak, the help provided by such a data mining system is valuable since it facilitates the decision making process.

REFERENCES

(2016). Retrieved from Greek School Network: www.sch.gr.

(2016). Retrieved from myschool: myschool.sch.gr.

Barnes, T., Desmarais, M., Romero, C., & Ventura, S. (2009). Educational Data Mining 2009. *2nd International Conference on Educational Data Mining, Proceedings.* Cordoba, Spain.

Bonner, M. R., Han, D., Nie, J., Rogerson, P., Vena, J. E., & Freudenheim, J. L. (2003). Positional accuracy of geocoded addresses in epidemiologic research. *Epidemiology, 14*(4), 408-411.

Burkom, H. S., Murphy, S., Coberly, J., & Hurt-Mullen, K. (2005, August 26). *Public Health Monitoring Tools for Multiple Data Streams*. Retrieved from CDC : Centers for Disease Control and Prevention: http://www.cdc.gov/Mmwr/preview/mmwrhtml/su5401a11.htm.

Castro, F., Vellido, A., Nebot, A., & Mugica, F. (2007). *Applying Data Mining Techniques to e-Learning Problems* (Studies in Computational Intelligence ed., Vol. 62). (T. T. Jain, Ed.) Springer-Verlag.

Cowen, J. D. (1988). *GIS versus CAD versus DBMS: what are the differences?* (Photogrammetric Engineering and Remote Sensing ed., Vol. 54).

Drummond, W. J. (1995). Address matching: GIS technology for mapping human activity patterns. *Journal of the American Planning Association, 61*(2), 240-251.

Garofalakis, J., Koskeris, A., & Vopi, A. (2007). An E-Government Application for Integrated, Multi-level Management of Large Scale resources of the Greek Primary and Secondary Education. *7th European Conference on e-Government (ECEG 2007).* The Netherlands.

Garofalakis, J., Koskeris, A., Michail, A.-T., Boufardea, E., & Oikonomou, F. (2011). An Information System to Collect and Analyze Data From Educational Units During Epidemy Spread Periods. *11th European Conference on e-Government (ECEG 2011).* Ljubljana, Slovenia.

Juan, A., Daradoumis, T., Faulin, J., & Xhafa, F. (2009). SAMOS: a model for monitoring students' and groups' activities in collaborative e-learning. *International Journal of Learning Technology, 4*(1-2), 53-72.

Kapageridis, I. (2006). *Introduction to Geographic Information Systems, Theory Notes.*

KEELPNO. (2010, February 03). *Weekly Epidemiological Report of Flu Virus, February 3rd 2010*. Retrieved 12 16, 2016, from Hellenic Center for Disease Control & Prevention: http://www.keelpno.gr/Portals/0/Αρχεία/Γρίπη και

Εποχική γρίπη/2003-2010/2009-2010/gripi_ebdo _20100203.pdf.

Mazza, R. (2009). *Introduction to Information Visualization.* Springer.

Mazza, R., & Milani, C. (2004). GISMO: a Graphical Interactive Student Monitoring Tool for Course Management Systems. *T.E.L.'04 Technology Enhanced Learning '04 International Conference*, (pp. 1-8). Milan.

Parker, D. H. (1988). *The Unique Qualities of a Geographic Information System: A Commentary* (Photogrammetric Engineering and Remote Sensing ed., Vol. 54).

Romero, C., & Ventura, S. (2010, December). Educational Data Mining: A Review of the State of the Art. *IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), 40*(6), 601-618.

Torres, L., Pina, V., & Acerete, B. (2005). *EGovernment developments on delivering public services among EU cities* (Government Information Quarterly ed., Vol. 22). (G.-G. J. Ramon, Z. Jing, & P.-C. Gabriel, Eds.) Elsevier.

Vine, M. F., Degnan, D., & Hanchett, C. (1998). Geographic information systems: their use in environmental epidemiologic research. *Journal of Environmental Health, 61*, 7-16.

West, D. M. (2002, September). *Global E-Government*. Retrieved from Inside Politics: http:// www.insidepolitics.org/egovt02int.html.

What is Epidemiology All About. (1999). *American Journal of Public Health Devotes August Issue to Epidemiology and Statistics*.