Data Mining for the Unique Identification of Patients in the National Healthcare Systems

D. G. Ramírez-Ríos[1], Laura P. Manotas Romero[1], Heyder Paez-Logreira[1], Luis Ramírez[1] and Yohany Andrés Jimenez Florez[2]

[1]Research Department, Fundación Centro de Investigación en Modelación Empresarial del Caribe, FCIMEC, Carrera 60 # 64-122, Barranquilla, Colombia

[2]Research Department, Logyca, Av El Dorado # 92-32, Bogotá, Colombia

Keywords: Data Mining, Databases, Secretary of Health, Duplicities, Healthcare System.

Abstract: This paper considers the application of data mining (DM) algorithms as a feasible and necessary strategy for optimal management of databases (DB) in the national healthcare systems. Specifically it deals with the management of multiple DB that consider patient's affiliation information, under the supervision of the authorities in healthcare, an issue that involves not only the issues of every citizen but also its integral right to be treated by any institution. We support the idea that the administrative part of the healthcare system should not obstruct the attention of the patient and a total efficiency must be guaranteed. We believe that DM algorithms are appropriate for this task and human intervention should be minimized. A case study was developed in Colombia that considered the multiple affiliations to DB and its integration to a unique DB managed by the District Health Secretary (DHS, which detected frauds and other type of duplicities. The mechanism used to approach this, indicates not only a significant reduction of manual intervention of the DB, but also allows the extraction of data for future analysis, supporting the patient's need for an efficient and integral health attention, as well as privacy of personal information registered.

1 INTRODUCTION

The institutions that provide healthcare services and the administrative entities that affiliate patients in the public sector are obligated to respond to the patient's needs. This implies the activities directly related to the health services provided for the patients and the administrative activities that allow the organization of their affiliations, removals and other services, which are done to provide an adequate and integral medical attention. According to the health policy, the authorities are in charge of regulating these activities and as part of this, the management of DB plays an important role in their daily activities.

The information stored in these DB must be organized, reliable, free of errors and duplicities, guaranteing its quality, and thus, assuring a correct and unique identification of the users and their full rights to the health services associated to his/her health plan. "This is why an implementation of unique health IDs are a requisite for the installation of politics and applications of the TICs in the sector" (Oviedo and Fernández, 2010).

The management of inconsistencies that can be detected inside the health information systems constitutes in a crucial aspect of the processing of data and is determinant over the benefits that a patient will or will not receive in terms of the services provided by the healthcare system. This is why a correct identification of users in a DB free of duplicities or multiple affiliations is strictly necessary and a responsibility of the entities in charge of the control and administration of the health sector.

One of the main problems identified in the management of DB is the correct identification of users and the organization of the data in the fields of the DB. This type of errors may cause problems during the formal affiliation of the patient (Esp and Ramírez, 2009). Furthermore and even more critical, a mistake in the affiliation attempts to the physical integrity of the patient, possibly aggravating its own health situation for a withdrawal or delay of the healthcare service required (McCoy, et al., 2013), not to mention the elevated costs involved in the administration of the health entities (McClellan, 2009).

G. Ramírez-Ríos D., P. Manotas Romero L., Paez-Logreira H., Ramírez L. and Andrés Jimenez Florez Y..
Data Mining for the Unique Identification of Patients in the National Healthcare Systems.
DOI: 10.5220/0005287302110217
In *Proceedings of the International Conference on Operations Research and Enterprise Systems* (ICORES-2015), pages 211-217
ISBN: 978-989-758-075-8

The entities in charge of the administration, attention and regulation of the subsidized healthcare services have a, even greater, responsibility of identifying the users of the health system, without making any mistakes. The information registered in the DB is necessary for accessing the government funds that covers the health plan of a subsidized patient. Thus, the local government entities of control and monitoring, periodically make revisions of the data registered in the information systems and check for inconsistencies reported with regards to subsidized users.

Given the actual situation, the information system plays an important role in guaranteeing the stability of the healthcare system. DBs are a key element in the administration of the information of users in the subsidized healthcare system. In order to guarantee that the system counts with complete and clean data (information free of errors), several DBs of users must be integrated, such as, the DB that contains the deceased, the new affiliations, the withdrawals, the transferred, among others. This integration must be done at a timely basis given that periodically there must be a report and the efficiency of the system must be maintained. DM is considered for this matter, given that it involves techniques and algorithms that allow correct and optimal management of DBs, as well as the use of information to gain knowledge over the population involved.

This research takes into account a point of view of the problem described above with respect to the correct identification and administration of inconsistencies in the registered data in the healthcare system. Particularly, this research identifies DM as an appropriate tool used for the timely detection of inconsistencies by many information systems. While some believe that DM is a robust and complex tool to be used for the detection of duplicities in the DB registrations in any information system, we believe it's completely necessary in order to get clean data and at the same time, obtain new knowledge from the data and a profound analysis of its behavior with respect to the abnormalities presented, that can become compelling to the overall quality of the system.

The paper is organized as follows: On section 2, a the state of art in DM applied to the Health Sector is given, supported by some applications with respect to duplicity detection on DBs; section 3, presents the case study developed, Unique Identification System for Users (SIUU) for the Health Sector, and guidelines of the solution are proposed; then, on section 4, Advantages and Disadvantages of DM in a SIUU, shows the importance of the DM for the detection of duplicities, patients' needs and the complexity of the solution; the last section presents concluding remarks and considerations to take into account when implementing the project.

2 DATA MINING APPLIED TO THE HEALTH SECTOR

A DB is a set of data that belong to the same context and are stored in a structural way for its further use (Date and Date, 1990). A DB provides institutions the access to information, in a way that it can be visualized, managed and updated, according to the access rights given (Batra, Parashar, Sachdeva, and Mehndiratta, 2013). With respect to the case study developed under this research, the DB identified as FOSYGA (MinSalud, 2014) is in charge of storing the Colombian healthcare information system with respect to the affiliation information. This DM provides access to sensitive information of the users registered in the system, which represents close to 91,69% of the entire Colombian population (DANE, 2013).

One of the most wearying activities to be done in terms of the administration of information is to keep the DB updated. In the Colombian healthcare subsidized system (RSS), local authorities must guarantee that the data updated is free of errors, since the payment given for the healthcare attention of a user that no longer belongs to the system is absorbed by the entities that offer the service and are not benefitting any other users. The identification of multiple registrations in this type of DB allows for a correct use of the government funds for healthcare services.

This same issue has been identified and approached in other countries, such as New Zealand, England, Spain, among others. In these countries, they have created a unique identification system for patients and have established some technological and legal frameworks in order to support and regulate the processes of affiliation and registration of patients in the system (Oviedo and Fernández, 2010).Yet, the problem is still present with or without the implementation of a unique identification system, given that the DB must be integrated and the data must be clean in order to use this information in the decision making process. DM has been approached to solve this issue, given that it gives the controlling entities the capacity to automatically classify and correct errors in the data.

DM has been important, not only in the organization and consolidation of DBs, it has also been of assistance with regards to statistics, mainly in the identification of vulnerable populations, types of treatments, geographic relationships and even the knowledge of the evolution of epidemics (Kaur and Wasan, 2006). Nevertheless, when DM is used for the regulation of DBs in the management of users, its impact is not too significant (Holzinger and Jurisica, 2014).

This leaves us the following questions: What if there existed such an integrated and complete DB of the users of the whole country, which contains from the registers of ID number to the number of deceased? How could it be exploited by using DM? What is the real impact in applying the complex algorithms found in DM into the management of user-based DB? The response to these questions leads us to consider several points of view. Some opinions may enhance the benefits, others the disadvantages of DM in several situations (Marcano Aular and Talavera, 2007) (Harrison, 2013) (Yucatan, 2014) (Harrison, 2013) (DeBariloche, 2014) (Pagliery, 2014).

There are multiple areas in which DM has been successfully implemented for the optimal management of DB (Hsu-Hao, 2012) in the health sector or user DB. Dávila Hernández and Sánchez Corales consider the concept “Clinical Decision Making Support Systems” (CDSS), which have proven to be fundamental in reducing the medical errors and improving the healthcare processes. A CDSS employs DM as a study method and is used for the classification of data, generating new knowledge from stored data. This research explains the contributions to the diagnosis of diseases using DM through the combination of two mathematical models. These models were applied to a case study on arterial hypertension and, as a result, behavior patterns were discovered with relation to factors that can raise the risk of having the disease (Davila Hernandez and Sanchez Corrales, 2012).

On the other hand, (Viveros, Nearhos, and Rothman , 1996) discuss the effectiveness of two techniques in DM that can be used to analyze and predict behavior patterns unknown to DBs that are registered by health insurance companies. The DB used were associated to pathology and general practitioners services. The techniques used in DM were association rules for pathology services and neural segmentation for the consolidation and evaluation of both DB integrated. The study demonstrates that DM algorithms can be used satisfactorily in huge data sets at a reasonable time and without employing too much computational effort. These results can be transformed into quantitative benefits and support decision making. Among the results shown, the study found an overpayment of more than $550.000 US per year, a figure that was not found in the conventional monitoring techniques.

It is possible to observe that DM is applied with much greater frequency in the follow up of treatments, diseases, patients and medicine in the health sector. With respect to the issues encountered in the administration of registered patient information of nationwide DBs that guarantees a correct and error free identification of users, DM is not too popular and it has been observed that for these cases, conventional algorithms are used for the detection of duplicities or other abnormalities.

Such a case is explained in (McCoy, et al., 2013), which evaluates the percentage of duplicities encountered in the Electronic Health Records (EHR) of five entities. This research establishes what is known as coincidence indicators, applied for the correct detection of duplicities, which can be easily adjusted to any entity. The algorithms employed have shown to be effective, given the increase in the amount of duplicities encountered.

On the other hand, DM and other advanced techniques used for the treatment of information stored in DBs have been applied for the detection of duplicities. In (Elmagarmid, Ipeirotis, and Verykios, 2007) a literature review is given and several methods used for solving the detection of duplicities are analyzed. For example, probabilistic approaches, automatic learning techniques and other variations, presenting the metrics and support tools in the application of systems for the detection of errors in DM.

3 UNIQUE IDENTIFICATION SYSTEM FOR USERS (SIUU) FOR THE HEALTH SECTOR, A CASE STUDY

A case study was implemented in the District Health Secretary (DHS), the dependency of the District Authorities, in charge of directing, coordinating and supervising the District’s Health System. It is in charge of providing healthcare services for the benefit of the community, such as, promotion, prevention of diseases, protection of the environment and health restoration. In order to fulfill all of its functions satisfactorily, the DHS requires integrated information of their users and statistics of the behavior of the health sector population in their municipality, in general.

The first step to obtaining the information of the health systems corresponds to the affiliation and registration of the users in the system. Counting with unique and reliable information of each patient is very important for guaranteeing a timely, efficient and economical attention in their healthcare plan.

The process of affiliation and registration of the patients is in charge of the Public Health Entities (EPS) that belong to the health system. These entities have the obligation of reporting the affiliations made to the National Health System and to the corresponding DHS. On the other hand, the DHS must supervise and control that the new affiliations are correct and unique for each patient.

In their process of supervision, the DHS compares the DB of the reported affiliations nationwide to other DBs, such as the National Registration, in order to verify data and identification of the users. These DBs are not integrated nor standardized, which implies that reported inconsistencies come in different formats, hindering the automatic integration of the DB.

This process is done at a monthly basis, where the EPS reports and sends documents weekly but the validation and verification takes longer, many times creating unsatisfied users and difficulties in the optimal response to situations presented.

The process of verification and validation of registered users face two delicate decisions: (1) removal of users from the RSS successfully registered in the system, because they are considered duplicates or invalid, creating a "false positive"; or (2) not detecting invalid users or belonging to another health system (including RSS, generating a duplication), and thus creating a "false negative". Both scenarios pose severe consequences, which implies having a user without health services or having a user affiliated twice, having a user who does not have the right to be benefitted or having one that does not even exist.

In order to effectively operate, a unique identification DB (SIIU), integrated and normalized, is proposed. For the structural model of the DB, data modelling and UML Class diagrams have been employed (Teorey, Lightstone, Nadeau, and Jagadish, 2011).With respect to the new integrated DB, it is possible to apply the techniques commonly used for the detection of duplicities, particularly through automatic learning. With the clean data and the duplicities identified, it is possible to apply DM for generating new knowledge from registered data, specifically the information concerning new affiliations, with regards to duplicities identified and common mistakes presented in the system. Other information concerning scope of healthcare attention and services provided, are also considered.

Figure 1 presents the actual and proposed mechanism for the affiliation of users and the detection of duplicities. The actual process begins with the EPS that sends the information both physically and digitally to the DHS, yet, the formats may vary among entities and the electronic media used for digital documents. The proposed mechanism is based on a server platform for the DHS in order to receive the requests digitally in a standardized document. The affiliation processes are to be supported by ITS and technologies for scanning documents automatically.

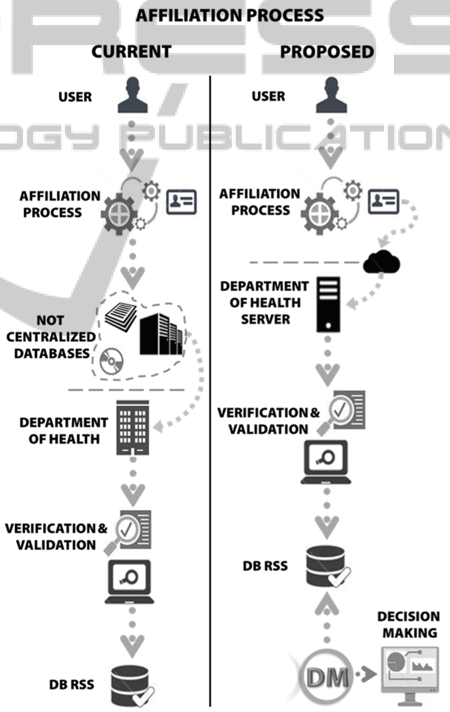

Figure 1: Affiliation process. Current and Proposed.

With the proposed mechanism, a few administrative steps are eliminated, yet the times of response and supervision processes are reduced significantly, making the process more efficient. With respect to the DBs, integrated and normalized, some DM techniques are applied in order to detect duplicities and other abnormalities (Elmagarmid, Ipeirotis, and Verykios, 2007).

The proposed mechanism considered in this case study is known to improve the times it takes in the processes of affiliation, validation and correction of registered data. Additional to this, valuable information that can be extracted from this DB through DM, such as, vulnerable population, most common emergency cases, types of attention required in the different age ranges, epidemic alerts, and entities with the most claims registered, among others.

4 ADVANTAGES AND DISADVANTAGES OF SIUU

Data mining is composed by a set of techniques widely recognized and applied in numerous fields. However, their use in certain applications must be evaluated from the perspective of the requirements, the ability of technological development and acquiring real customer needs, aside from the fashions and preferences based on developer experience.

From an initial review "DM is a technique that optimizes and improves the effectiveness in detecting duplicate records in databases." The detection of duplicate records is one of the simplest processes in Data Cleansing. Cleaning records in a database is studied in conjunction with data mining and other areas.

Comparing the records, one by one, is the most reliable duplication detection process in a database. However, this technique is time consuming, demands lots of processing resources and is conditioned on the number of records to evaluate (Wai Lup, Mong Li, and Tok Wang, 2001).

Other techniques have been proposed, using algorithms such as “nearest neighbor”, to reduce the consumption of time and resources in duplication detection algorithms. Another technique is to have a stack of records previously detected and prioritize these to be the first to be evaluated.

Lower complexity algorithms of Data Mining and Data Cleansing are applied in duplicate detection such as Soundex, assessing a 95.99% effective records (Elmagarmid, Ipeirotis, and Verykios, 2007). These algorithms or methods are suitable for detecting duplicities on individual fields. Nevertheless, the detection record consisting of multiple fields is a more complex problem, which requires the application of probabilistic approaches and supervised machine learning techniques, used in DM.

A second opinion suggests that "DM is a very complex and robust technique to be applied to a single user registration process." DM is not an easy task and consumes a lot of human and equipment resources. DM implementation involves the acquisition of query and analysis tools and training the users (Xintong, Hongzhi, Song, and Hong, 2014).

The challenge of DM, but also one of its advantages, is to be a framework that integrates multiple approaches from different disciplines and knowledge areas (Bellazzi and Zupan, 2008). The application of DM in a specific domain problem requires that developers are not only experts in DM but also acquire a considerable level of knowledge about the problem itself. Application of Data Mining in a specific field require a proper analysis of the problem domain and modeling solution (Shu-Hsien, Pei-Hui, and Pei-Yuan, 2012) to establish an appropriate methodology for the problem, for example, detection of records and duplication in the Health Sector.

When analyzing the information related to people and the method of implementation of DM for Data Cleansing, some negative aspects may appear, which are classified by several authors in four general key factors claves (Han and Gao, 2008):Security: Although entities can manage large amounts of data that contain personal and confidential information of the users, in occasions, there are no mechanisms that prevent the loss and/or stealing of data, generating a risk to the security of the users registered.

1. Privacy: DM requires data to be exposed to the processes applied, so it is necessary to accompany this information with the appropriate security techniques and encryption protocols.
2. Accuracy: Sometimes an error in the data processing could generate a huge problem if the information is interpreted or processed incorrectly.
3. Complexity: Huge investment for processing information.

From these two points of view, the application of DM in the District’s DB for a system of duplication detection must be evaluated. Assessing customer needs and the relevance of the proposed solution.

For this case study, sessions were made for the capture of requirements and modeling the problem domain. The group was formed by researchers and stakeholders involved from the perspective of the business. This process is recorded in class diagrams and requirement diagrams for the domain model by using UML.

From these diagrams, the identification of processes that can be automated or implemented in software was performed, the selection of features or cases of use that guide the development process, defining an architecture for the project and therefore the implementation of the solution.

5 CONCLUSIONS AND FURTHER RESEARCH DIRECTIONS

In this paper, the specific problem of detecting duplicate records and Data Cleansing on the User Record System of the DHS is presented. In addition, positions are presented against the application of DM to solve this problem, emphasizing the importance of using software development methodologies, modeling languages like UML and analysis requirements for making the right decision of software architecture and techniques to be applied in the solution.

It is evident that the detection of duplicate records is a problem of special attention for the DHS, affecting the available economic, quality assurance and patient care time.

A web architecture is proposed to streamline the registration process for new members of the healthcare system and the digitalization of the documents that must be delivered, reducing the use of paper and human intervention or manipulation of information.

There are two opinions on the application of DM to the problem of detecting duplicate records in DBs. In the first case, there are advantages and benefits that this technique can contribute to the problem and secondly, the complexity and relevance that results from its application.

In the case study it was determined the application of algorithms for the detection of duplicities, as Soundex, with optimization and clustering techniques to reduce the execution and detection times. Additionally, DM is used to analyze the results in duplicities in order to prevent further duplication or fraud attempts entering invalid records system records.

Mining on the results obtained can also generate constructive knowledge of DBs, as advanced statistics and forecasts hedging, investment plans in the healthcare system, affiliation programs to the health care system, among others.

As future research directions it is important to take into account different DM algorithms and compare their results in this specific field of application to evaluate their performance, effectiveness and appropriateness of these, enabling support implementation decisions to the solution presented.

ACKNOWLEDGEMENTS

This research was supported by the Colombian Institute for Development in Science and Technology Francisco José de Caldas” (COLCIENCIAS) through the project titled “Desarrollo de un sistema de información para la gestión de usuarios en la Secretaria de Salud del Distrito de Barranquilla” No. 3351-604-37776.

REFERENCES

Batra, S., Parashar , H., Sachdeva, S., and Mehndiratta, P. (2013). Applying data mining techniques to standardized electronic health records for decision support. *ieeexplore*, 510-515.

Bellazzi, R., and Zupan, B. (2008). Predictive data mining in clinical medicine: current issues and guidelines. *International journal of medical informatics, 77*(2), 81-97.

DANE. (2013). *Departamento Administrativo Nacional de Estadística.* Obtenido de https://www.dane.gov.co/index.php/estadisticas-sociales/encuesta-longitudinal-de-proteccion-social.

Date, C., and Date, C. (1990). An introduction to database systems. *Addison-wesley Reading, MA*.

Davila Hernandez , F., and Sanchez Corrales, Y. (2012). Técnicas de minería de datos aplicadas al diagnóstico de entidades clínicas. *Revista Cubana de Informática Médica*, 174-183.

DeBariloche. (09 de 15 de 2014). *Portal Rio negro*. Obtenido de Buscan mejorar la Tish con la base de datos de AFIP: http://www.rionegro.com.ar/diario/buscan-mejorar-la-tish-con-la-base-de-datos-de-afip-4365198-9862-nota_cordillera.aspx.

Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *Knowledge and Data Engineering, 19*(1), 1-16.

Esp, I., and Ramírez, C. (2009). Hacia una metodología para la selección de técnicas de depuración de datos. *Rev. Av. En Sist. E Informática 6*.

Han, J., and Gao, J. (2008). Research challenges for data mining in science and engineering. *Next Generation of Data Mining.* . Chapman and Hall.

Harrison, T. (07 de 2013). *PRWeb.* Obtenido de http://www.prweb.com/releases/2014/08/prweb12084827.htm.

Holzinger, A., and Jurisica, I. (2014). Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions. *Interactive Knowledge Discovery and Data Mining in Biomedical Informatics*, 1-18.

Hsu-Hao, T. (2012). Global data mining: An empirical study of current trends, future forecasts and technology diffusions. *Expert Systems with Applications, 39*(9), 8172-8181. Obtenido de http://www.sciencedirect.com/science/article/pii/S0957417412001704.

Kaur, H., and Wasan, S. K. (2006). Empirical Study on Applications of Data Mining Techniques in Healthcare. *Journal of Computer Science*, 194-200.

Marcano Aular, Y. J., and Talavera , R. (2007). Minería de Datos como soporte a la toma de decisiones empresariales. *Opción*, 10-118.

McClellan, M. A. (2009). Duplicate medical records: a survey of twin cities healthcare organizations. *AMIA Annual Symposium Proceedings*, 421.

McCoy, A. B., Wright, A., Kahn, M. G., Shapiro, J. S., Bernstam, E. V., and Sittig, D. F. (2013). Matching identifiers in electronic health records: implications for duplicate records and patient safety. *BMJ quality and safety, 22*(3), 219-224.

MinSalud. (09 de October de 2014). *FOSYGA*. Obtenido de http://www.fosyga.gov.co/

Oviedo, E., and Fernández, A. (2010). Tecnologías de la información y la comunicación en el sector salud: oportunidades y desafíos para reducir inequidades en América Latina y el Caribe. *CEPAL*.

Pagliery, J. (09 de 09 de 2014). *Home Depot confirms months-long hack*. Obtenido de http://money.cnn.com/2014/09/08/technology/security/home-depot-breach/

Shu-Hsien, L., Pei-Hui, C., and Pei-Yuan, H. (2012). Data mining techniques and applications – A decade review from 2000 to 2011. *Expert Systems with Applications, 39*(12), 11303-11311.

Teorey, T. J., Lightstone, S. S., Nadeau, T., and Jagadish, H. V. (2011). Database modeling and design: logical design. *Elsevier*.

Viveros, M., Nearhos, J., and Rothman , M. (1996). Applying Data Mining Techniques to a Health Insurance Information System. *Proceedings of the 22th International Conference on Very Large Data Bases* (págs. 286-294). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Wai Lup, L., Mong Li, L., and Tok Wang, L. (2001). A knowledge-based approach for duplicate elimination in data cleaning. *Information Systems, 26*(8), 585-606.

Washington, A. P. (10 de 23 de 2014). *newspaper the guardian*. Obtenido de http://www.theguardian.com/technology/2014/aug/23/homeland-security-25000-employees-hacked.

Xintong, G., Hongzhi, W., Song, Y., and Hong, G. (2014). Brief survey of crowdsourcing for data mining. *Expert Systems with Applications, 41*(17), 7987-7994. Obtenido de http://www.sciencedirect.com/science/article/pii/S0957417414003984.

Yucatan, D. d. (17 de 09 de 2014). *Diario de Yucatan*. Obtenido de http://yucatan.com.mx/merida/policia/recibe-reconocimiento-fge-por-sistema-forense.