

A Multidimensional Data Modeling of the SEER Database from the USA National Cancer Institute

Heidy M. Marin-Castro, Jose Torres-Jimenez and Diana I. Escalona-Vargas
Center of Research and Advanced Studies of the National Polytechnic Institute
Information Technology Laboratory
Scientific and Technological Park of Tamaulipas TECNOTAM, km 5.5
Cd. Victoria, Tamaulipas
{hmarin,jtj,descalona}@tamps.cinvestav.mx

Abstract—Nowadays, one of the main challenges in computer science is to process the large amount of data available in diverse data sources, such as databases or files, in order to find useful information. For this purpose, it is required specialized tools that process raw data in a smart way to discover knowledge. In this paper, we present the design of a data warehouse and a tool called TDR (Tool Drill-Roll) that allow to discover knowledge from the database SEER (Surveillance, Epidemiology, and End Results) from the Cancer Institute in the United States of America, which has more than five million of records. The data warehouse is designed using a multidimensional approach and the TDR tool allows to exploit interesting information from SEER using drill-down and roll-up (two operators of On line Analytical Processing (OLAP)). The data warehouse can be seen at many levels of granularity. Our developed TDR tool allows knowing the statistics of the incidence, mortality and survival of patients with cancer along of years and extract useful information related to this disease that could be used to establish a relation between certain characteristics of patients that has an specific type of cancer. The knowledge discovered by our TDR tool could be of interest for government, health care institutes or research community for decision making. The main contribution of this paper is the discovery of new knowledge from the SEER database. The methodology used to design the data warehouse and the TDR tool could be applied to others domains with minimal changes.

Keywords-data warehouse; OLAP; drill-down; roll-up; cancer-database.

I. INTRODUCTION

Cancer is a disease fundamentally characterized by uncontrolled growth of malignant cells (known as carcinogenic or cancerous). The associated terminology with the issue of cancer is between benign and malignant tumors. A tumor or neoplasm is any abnormal growth of cell, which may be either benign or malignant. A benign tumor remains confined to its original location. It neither invades surrounding normal tissue nor spreads to distant body sites. In contrast, a malignant tumor is capable both of invading adjacent normal tissue and of spreading to others tissues and organs [1]. There are more than 100 different types of cancer. Most cancers are named for the organ or type of cell in which they start [2]. According to a report presented in 2010 of the

U.S Population [3] approximately 789,620 men and 739,940 women will be diagnosed with cancer. Also, it is estimated that 569,490 men and women will die due to cancer in 2010 considering all cities in the United States of America.

In this paper, we present the construction of a data warehouse and a software tool called TDR to exploit interesting information related with records of patients with cancer in the United States of America using OLAP. We use information provided by the SEER (Surveillance, Epidemiology, and End Results) database which contain records of patients with different types of cancer from 1978 to 2007. Our design is based on a multidimensional approach for building the data warehouse. We created a six dimensional model that allows to uncover and grouping information not yet discovered with different levels of granularity. The multidimensional model was built in SQL language. To query the data warehouse we implemented a tool called TDR (Tool Drill-Roll) that incorporates the operators drill-down, roll-up and slice-dice. This operators acts on the defined hierarchies for exploring and visualizing the information at different levels of granularity. This ability makes our tool very useful to discover new knowledge from the SEER database that could be used by National Cancer Institute for decision making. TDR is a friendly tool that present information in tabular and graphical form making its use more easy to the end user. TDR has a browser of information which displays 3D graphs of the results of OLAP operators. Our design methodology for the data warehouse could be applied to different domains or databases with minimal changes.

Unlike well known tools for multidimensional analysis as Business Object [4] and Talend [5], the TDR tool presented in this work allows greater flexibility to add ad-hoc analysis on the data, and the possibility of calling proper functions at the time the data is being manipulated visually. In addition, our tool allows visualizing the discovered information displaying the recovered data in a 3D bar chart.

The rest of this document is organized as follows: next section presents an overview of the literature, describing general and related concepts and techniques. Section 3 describes the design process of the proposed data warehouse,

explaining the dimensional approach and how hierarchies were defined. Section 4 describes the six-dimensional data cube designed in this work. Section 5 presents the architecture of the TDR tool. In Section 6 are presented the results achieved in this work and a discussion of them. Finally, Section 7 presents the conclusions of this work.

II. LITERATURE REVIEW

The concept of data warehouse was defined by W. H Inmon [6] as a subject-oriented, integrated, time variant and non-volatile collection of data in support of managements decision making process. Another definition states that a data warehouse is a collection of large amount of data or a repository (collection of resources that can be accessed to retrieve information) of organizations electronically stored data, useful to generate reports [7]. The data warehouse paradigm provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decisions. Data warehouse systems are valuable tools in today competitive, fast-evolving world. The data warehouse management is a relational database that contains the data that is collected from a server that is a data collection target. This data is used to generate the reports for the System Data collection sets, and can also be used to create custom reports.

The drill-down, roll-up and slice-dice operators are included in the design of a data warehouse manager. These operators allow the user to view the data at differing degrees of summarization. The first operator, drill-down builds a view to go from a general level of granularity to an specific level. The roll-up operator works in inverse sense to drill-down, roll-up builds a view to go from an specific level to a general level.

The data warehouse paradigm has been used in many business contexts. Recently, it is being used in the biomedical field such as biological science on clinical and genomic data [8]. The objective is to improve the capabilities of the On-line Analytical Processing (OLAP) to make effective medical decisions [9]. In the literature, there are few works related to a multidimensional approach for processing data in the domain of cancer disease. In [10] Wahet *al.* developed a complete multidimensional model design of a data warehousing for a Lymphoma cancer used the relevant information from the Internet. They proposed a six dimensional model. However, in this work the design of hierarchies is not considered as we do in our work.

III. DATA WAREHOUSE ARCHITECTURE

A. Database description

The study presented in this work uses the new version of SEER [2] database during the years from 1973 to 2007 with 5,178,804 records covering to eight cancer types (breast, colon and rectum, respiratory, female genital, urinary, male genital, lymphoma in all sites, leukemia, and digestive). We

#	Item name	Positions	Length
1	Patient ID number	01-08	8
2	Registry ID	09-18	10
3	Marital Status at DX	19-19	1
4	Race/Ethnicity	20-21	2
7	Sex	24-24	1
8	Age at diagnosis	25-27	3
12	Month of diagnosis	37-38	2
13	Year of diagnosis	39-42	4
20	Grade	58-58	1
21	Diagnostic Confirmation	59-59	1
87	Age Recode < 1 Year olds	183-184	2
107	SEER Summary Stage 1977	231-231	1
108	SEER Summary Stage 2000	232-232	1
115	Vital Status recode	255-255	1

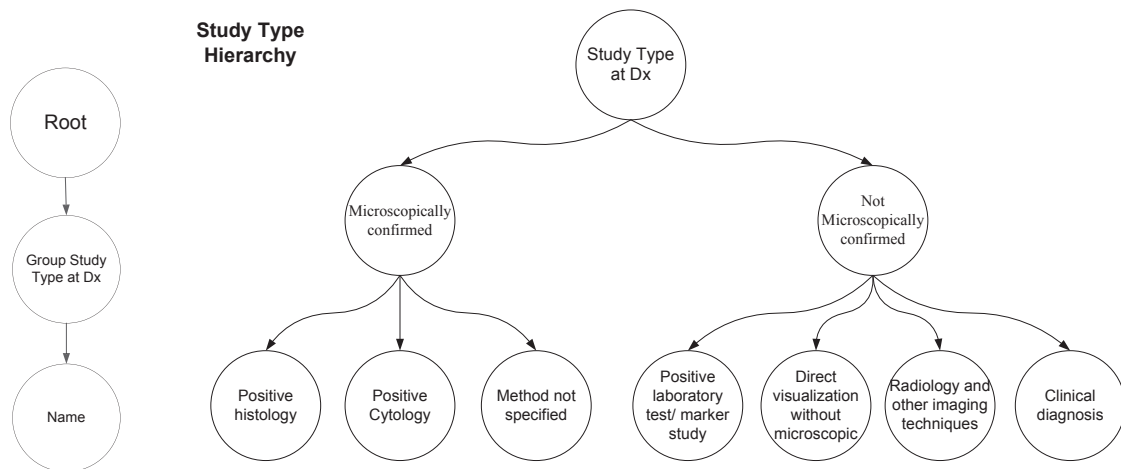
TABLE I
ATTRIBUTES SELECTED FROM MULTIDIMENSIONAL MODEL.

organized the information of this database in dimensional tables to exploit information of patients with any type of cancer. Our study uses the attributes show in Table I of Computer Record Format (CRF) of SEER data fields.

B. Facts and Dimensional Hierarchies

The number of patients with cancer has been increasing in recent years. It could be interesting to know the statistics of the incidence, mortality or survival along of years and extract useful information related of the disease and establish a relation between certain attributes of patients that has a type of cancer. Therefore, the most important characteristics of a data warehouse is the fact table that allows to quantify the number of incidences of cancer recorded. Our study consists in the quantification of medical cancer events that happened on certain dates, we can view how many patients are alive or dead with an specific type of cancer and with certain characteristics related to sex, group of age at diagnosis (child, young, adult, senior), marital status. The design of the data warehouse is based on a dimensional approach.

In this approach, the design is guided by the definition of hierarchies that allows to exploit data at different granularity levels. These hierarchies are designed from the tables in SEER database according to the hierarchical relation among the fields in such tables. In this work, these hierarchies are expressed as a direct acyclic graph. One example of these hierarchies is shown in Figure 1. In this figure, to the left there is the hierarchy of the fields in the corresponding dimension. Each node in the hierarchy is referred as a *category*. To the right of the figure is the same hierarchy but in terms of the possible *values* for each category. The possible values for category *group study type* are *microscopical and non-microscopical confirmed*. In the next level of this hierarchy is the category *name* with values *positive histology, positive cytology, etc*. Hierarchies, categories and values play an important role in the design and implementation of the proposed data warehouse in this work.

Figure 1. Hierarchy model from dimension *type of study*

All information in tables of SEER database are related by means of a table of facts called *medical_facts*. This table stores facts about patients such as if they are alive or dead with due to specific type of cancer and with certain characteristics related of sex, age group of diagnosis (child, young, adult, senior), marital status among others characteristics.

This characteristics are called *variables*, and they are variables of interest tracked by the TDR tool. The diagram of the table of facts allows to build reports that answer questions such as:

- The number of patients of black race diagnosed with breast cancer from 1990 to 2000.
- The number of women died due to mama cancer.
- The predominant marital status of patients having colon rectal cancer.

IV. MULTIDIMENSIONAL DATA CUBE

We created OLAP cubes from a data model in order to obtain answers to questions such as the ones listed at the end of section III-B. An OLAP cube is not strictly a cuboid [11], it is the name given to the process of linking data from the different dimensions. In this work the data cube is created using four out of seven available dimensions.

Figure 2 shows a cuboid using three dimensions. To the right of this figure it is shown the four ways of data can be explored from this cube by fixing categories and values at each dimension. In this work we use a data cube defined by four dimensions (X, Y, Z, W) instead three. This data cube is commonly known as a *teserac*. For dimension four (Z), we fix one from all possible values for this dimension. Then the associated cuboid given by (X, Y, Z) is processed

by OLAP operators in order to get the four possible view of data.

A. Refinement OLAP operators or querying manipulation

The main advantage when building a data warehouse is the use of operators for facilitating the aggregation (consolidation) or the disaggregation (division) of data. The operator *roll-up* (aggregation) allows to eliminate a grouping criterion. The *drill-down* operator (disaggregation) allows to introduce a new grouping criterion. These operators allow querying the data source that can not be done using traditional operations like selections, projections, concatenations or groupings.

Drill-down and roll-up OLAP operators make extensive processing of hierarchies described in Section III-B. When a dimension is selected in the *teserac*, all records of patients in the data source are considered for querying (no grouping restrictions are given). The hierarchies associated to the selected dimensions are reset to the root node. From this point, an specialized query can be done by selecting one of the reached categories from the root node in one of the hierarchies. The result is the selection of an *slice*, a subset of data such as the ones shown in Figure 2. Several selected slices form a *dice*. For example, when a category is selected, all values for this category define several slices from the data cube, that is, they define a *dice*.

If we want to descend in a hierarchy, we need to fix a value for the current category, which means to reduce the data set only to that records meeting the condition *category = value*. Now, for the new data subset we can repeat the same process in case more descendants exist in the hierarchy. Each time we perform a drill down operation a new condition given by the expression *category=value* must

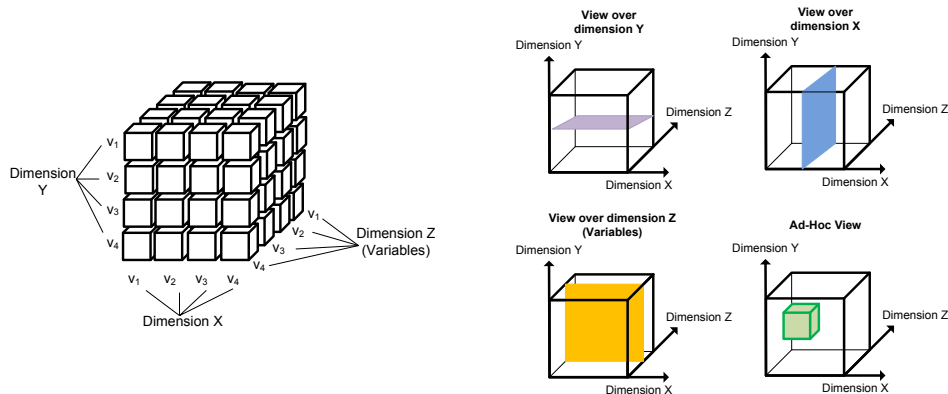


Figure 2. Cuboid and four ways to explore data

be satisfied, together with the previous conditions. So, the drill-down operator defines a *path* that specifies a set of conditions that group data for specialized querying. Drill down operator can be applied over any dimension at any time. The data grouped by the drill down operator is dis grouped by the roll up operator. This means that we need to go back one condition in the current built path. Doing this causes the immediately previous data set be recovered and used for querying.

Both drill down and roll up operators allow to obtain a measure of facts, with restrictions given by the conditions imposed over dimensions. The number of possible data subsets that can be explored is given by the numbers of paths that can be formed from the hierarchies, from the root node to each leaf in the graph.

V. A GRAPHICAL USER INTERFACE FOR THE MULTIDIMENSIONAL MODEL

We built a software tool called TDR to get data out for users by applying the drill down and roll operators up over the SEER database. However, this tool allows operate over any other database with minimal changes.

All information about the dimensions, facts and hierarchies is stored in a database, so, this tool is scalable to support new dimensions or hierarchies. In the reports we focus on a single variable from the table of facts, like “survival” or “mortality”. This is also possible for the user to select the tracked variable.

Users can decide how to explore data in the SEER database by selecting any of the available dimensions presented in Section 2.3. For each dimension, the user can go down across the hierarchy associated to that dimension for a more specialized query. If required, the user can go back the hierarchy or select a different dimension at any time. Always a dimension is selected, the associated hierarchy is recovered and the navigation starts form the root node. Drill down is the main operator for doing specialized queries to the database. At the beginning, suppose that one of the

selected dimensions is “age”. The hierarchy associated to this dimension indicates that the first criterion is “year of diagnosis”. Then, the tool presents all possible values in the facts for this given current criterion. If the user wants to descend in this hierarchy, he performs a drill down operator by selecting one of the possibles values of this current criterion. The hierarchy associated to dimension “age” indicates that the next criterion is “month of diagnosis”, so now we see in the output all possible values for this now current criterion. The user can proceed in the same way until no more descendants exist in the hierarchy. A roll up operation on a hierarchy causes the grouping done by the previous drill down operation be dissolved. Slice and dice operations are implemented by selecting specific values for current categories selected in dimension *Z* and dimension four. Dimension *Z* is assigned to the tracked variables in the fact table. The fourth dimension is traversed by its corresponding hierarchy. In this case all possibles values for the current criterion in the hierarchy is not shown in the table of results due to this table will be very large in the browser. Instead, just a single value is considered at a time, which ca be selected from a combo box.

The software tool uses a plotting library for 3D graphics that can be downloaded from [12]. All tables needed for operating like dimensions and hierarchies are stored in a database in MySQL. The tool uses the dimensions: *age-dimension*, *cancer-dimension*, *marital-status-dimension*, *race-ethnicity-dimension*, *sex-dimension*, *study-dimension*.

Our developed tool uses four dimensions that can be selected among the ones previously registered. From these dimensions the user can determine the search criterion for answering a given question for example “*How many women died because of a not solid cancer type in the year 1990?*”. In this case, the user must select the dimensions *sex*, *cancer-type* and *age-diagnosis*. If user wanted specialize the above query, he only needs to do a drill down operation on any of the available dimensions, for example, specializing the above query to the month September in the same year of

1990.

VI. RESULTS

In this study, we used the TDR tool to obtain new and relevant knowledge about cases of cancer in patients that live or die in the United States between the year 1973 to 2007. In the first analysis we obtained the rate of mortality and survival of women with mama cancer. The results shown 104677 widowed women died due to mama cancer and 83448 survived. For this case, the rate of mortality and survival is shown in the equations 1, 2 respectively. For the case of single women, the rate of mortality of single women is 31.67% and the rate of survival is 68.33%. These results show that the mortality in widowed women is higher than in single women. However, the survival rate for married women is 71%, being the highest. It seem to be that the marital status has an important effect asociated with the mortality and the risk of mortality of widowed women over married women.

$$mortality_rate = \frac{104677 * 100}{188,125} = 55.64 \quad (1)$$

$$survival_rate = \frac{83448 * 100}{188,125} = 44.36 \quad (2)$$

In US, rectal colon cancer is the fourth cancer with more incidence. It is more common in old people (around 50s) and risk increasing with age. In our study we observe that this cancer type affects in equal number to women and men. This keeps true for patients that survive to this cancer.

Another interesting uncovered fact is that most of the realized studies for detection of cancer were confirmed microscopically. It seems to be that patients rely on this technique. We observe that the number of died patients due to mama cancer diagnosed microscopically is higher than the number of survived patients. In this case, the number of patients died due to not solid cancer (leukemia) diagnosed using any kind of study (not a microscopically study) is around 10287. The number of cases confirmed microscopically were 272913 and the case confirmed not microscopically were 9673.

VII. CONCLUSION

This work presented the development of a data warehouse based on the SEER database to find relevant and useful knowledge in data. We developed a tool called TDR that makes extensively use of operators drill-down, roll-up, and slice-dice to explore data at different granularity levels. We found interesting information on the SEER database such as the mortality and survival rate of patients for different types of cancer. The development of data warehouse required of some considerations. First, data cleaning played the most critical role in the data warehouse development. It is a time consuming task that must be handled thoroughly to avoid no useful data. Second, the structure of the data warehouse is designed in a flexible way allowing further extensions

and adaptation to other domains databases. Finally, the dimensional approach used for the construction of the data warehouse allowed defined the granularity of queries at different levels.

The results presented in this work can be extended or new knowledge could be explored using our TDR tool. The end user has fully freedom to form questions and getting the answers using our tool. For example, “How many women were diagnosed with breast cancer in 1980?”.

ACKNOWLEDGMENTS

This research was partially funded by the following projects: CONACyT 58554-Cálculo de Covering Arrays, 51623-Fondo Mixto CONACyT y Gobierno del Estado de Tamaulipas.

REFERENCES

- [1] G. M. Cooper, *The Cancer Book: A Guide to Understanding the Causes, Prevention, and Treatment of Cancer*. London,UK: Printed in the United States of America, 1993.
- [2] National Cancer Institute, “Surveillance Eoidemiology and End Results,” <http://seer.cancer.gov>.
- [3] National Cancer Institute SEER, “NCI Fact Sheets,” <http://seer.cancer.gov/cancertopics/factsheet>.
- [4] SAP, Company, “Soluciones SAP para el almacenamiento de datos ,” <http://www.sap.com/mexico/solutions/sapbusinessobjects/data-warehousing/index.epx>.
- [5] www.datawarehouse4u.info, “Talend Open Studio ,” <http://datawarehouse4u.info/Talend-Open-Studio.html>.
- [6] W. H. Inmon and R. D. Hackathorn, *Using the Data Warehouse*. New York: Wiley-QED, 1994.
- [7] P. Vassiliadis, C. Quix, Y. Vassiliou, and M. Jarke, “Data Warehouse Process Management,” *Information Systems*, vol. 26, pp. 205–236, 2001.
- [8] L. Wang, A. Zhang, and M. Ramanathan, “Biostar models of clinical and genomic data for biomedical data warehouse design,” *Int. J. Bioinformatics Res. Appl.*, vol. 1, pp. 63–80, April 2005. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1356096.1356101>
- [9] S. Palaniappan and C. S. Ling, “Clinical Decision Support Using OLAP With Data Mining,” *International Journal of Computer Science and Network Security*, vol. 8, no. 9, 2008.
- [10] T. Y. Wah and O. S. Sim, “Development of a data warehouse for lymphoma cancer diagnosis and treatment decision support,” *WSEAS Trans. Info. Sci. and App.*, vol. 6, pp. 530–543, March 2009. [Online]. Available: <http://portal.acm.org/citation.cfm?id=1553642.1553661>
- [11] R. Kimball and J. Caserta, *The Data WarehouseETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data e*. Canada: Wiley, 2004.
- [12] L. Tautenhahn, “SVG-VML-3D 1.3 Graphic Library,” available at <http://www.lutano.net/svgvml3d/index.html>.