Biomedical Data Warehouses

Jérôme Darmont
University of Lyon 2, France

Emerson Olivier
University of Lyon 2, France

INTRODUCTION

With the growing use of new technologies, health care nowadays is undergoing significant changes. The development of electronic health records will indeed help enforce personalized, lifetime health care and presymptomatic treatment, as well as various analyses over populations of patients. Information-based medicine has to exploit medical decision-support systems and requires the analysis of various, heterogeneous data, such as patient records, medical images, biological analysis results, and so forth (Saad, 2004).

Data warehousing technologies (Inmon, 2002; Kimball & Ross, 2002) are now considered mature and can form the base of such a decision-support system. Although data warehousing primarily allows the analysis of numerical data, its underlying concepts remain valid for what we term *complex data*. To summarize, data may be qualified as complex if they are (Darmont, Boussaïd, Ralaivao & Aouiche, 2005):

- **Multiformat:** Represented in various formats (databases, texts, images, sounds, videos) and/or
- **Multistructure:** Diversely structured (relational databases, XML document repositories, file collection) and/or
- **Multisource:** Originating from several different sources (distributed databases, the Web) and/or
- **Multimodal:** Described through several channels or points of view (radiographies and audio diagnosis of a physician, data expressed in different scales or languages) and/or
- **Multiversion:** Changing in terms of definition or value (temporal databases, periodical surveys with evolving items)

In this context, the warehouse measures (although not necessarily numerical) remain the indicators for analysis, and analysis is still performed following different perspectives represented by dimensions. Large data volumes and their dating are other arguments in favor of this approach (Darmont, Boussaïd, Bentayeb, Rabaseda & Zellouf, 2003). Data warehousing also can support various types of analysis, such as statistical reporting, online analysis (OLAP), and data mining.

The aim of this chapter is to present an overview of the existing data warehouses for biomedical data and to discuss the issues and future trends in biomedical data warehousing. We illustrate this topic by presenting the design of an innovative, complex data warehouse for personal, anticipative medicine.

BACKGROUND

The first family of medical data warehouses we identify are repositories tailored for supporting data mining (Miquel & Tchounikine, 2002; Prather, Lobach, Goodwin, Hales, Hage & Hammond, 1997; Sun, Huang, Horng, Huang & Tsou, 2004; Tchounikine, Miquel & Flory, 2001). However, since data mining techniques take attribute-value tables as input, some of these warehouses (Prather et al., 1997; Sun et al., 2004) do not bear the multidimensional, starlike architecture that is typical in data warehouses. This modeling choice precludes OLAP navigation and is not very evolutionary; new analysis axes, or dimensions, cannot be easily plugged into the warehouse.

The most "canonical" medical data warehouse among these proposals is a cardiology data warehouse (Miquel & Tchounikine, 2002; Tchounikine et al., 2001). Its aim is to ease medical data mining by integrating data and processes into a single warehouse. However, raw sensor data (e.g., electrocardiograms) are stored separately from multidimensional data (e.g., patient identity, therapeutic data), while it might be interesting to integrate them all.

A second family is constituted of biological data warehouses that focus on molecular biology and genetics (Engström & Asthorsso, 2003; Eriksson & Tsuritani, 2003; Schönbach, Kowalski-Saunders & Brusic, 2000; Shah, Huang, Xu, Yuen, Ling & Ouellette, 2005; Sun et al., 2004), and bear interesting characteristics. For instance, some of them include metadata and ontologies from various public sources such as RefSeq or Medline (Engström & Asthorsso, 2003; Shah et al., 2005). The incremental maintenance and evolution of the warehouse is also addressed (Engström & Asthorsso, 2003). However, the particular focus of these approaches makes them inappropriate to more general needs, which may be both different and much more diversified.

Eventually, Boussaïd, Ben Messaoud, Choquet, and Anthoard (2006) recently proposed an eXtended Markup Language (XML) based methodology named X-Warehousing for warehousing complex data. This approach has been applied on a corpus of patient records extracted from the Digital Database for Screening Mammography[1]. The warehouse is a collection of XML documents representing OLAP facts that describe suspect areas in mammographies. It aims at breast cancer computer-aided diagnosis.

A COMPLEX DATA WAREHOUSE FOR PERSONALIZED, ANTICIPATIVE MEDICINE

Context and Motivation

Dr. Jean-Marcel Ferret, former physician of the French national soccer team, is the promoter of the personalized and anticipative medicine project. His aim is to extend results and empirical advances achieved for high-level athletes (not only soccer players) to other populations and to make the analyzed subjects the managers of their own health capital by issuing recommendations regarding, for example, life style, nutrition, or physical activity. This is meant to support personalized, anticipative medicine. In order to achieve personalized, lifetime health care and presymptomatic treatment, a decision-support system must allow transverse analyses of given populations of patients and the storage of global medical data such as biometrical, biological, cardiovascular, clinical, and psychological data. It also must be evolutionary in order to take into account future advances in medical research. More precisely, such a system must be able to store complex medical data and allow quite different kinds of analyses to support the following:

1. Personalized and anticipative medicine (in opposition to curative medicine) for well-identified patients;
2. Broadband statistical studies over given populations of patients.

We selected a data warehousing approach to answer this need. A data warehouse can indeed support statistical reporting and cross-analyses along several dimensions. Furthermore, dated personal data can be stored to propose full patient records to physician users. However, data complexity must be handled. For instance, multimedia documents such as echocardiograms must be stored and explicitly related to more classical medical data such as the corresponding patient or diagnoses. Users must be able to display and exploit such relationships, either manually (which is currently the case) or automatically (here, we anticipate the advances of multimedia mining and the development of advanced OLAP operators). The existing research proposals from the literature do not fulfill this requirement. In particular, Tchounikine, et al. (2001) store raw sensor data separately from multidimensional data, only as an archive. Finally, Dr. Ferret had quite an immediate need for an operational, efficient system. Hence, we chose to rely on the efficacy of a relational implementation. Although XML warehousing is particularly appropriate to complex, medical data, the performance of native XML Database Management Systems (DBMSs) is indeed currently not satisfactory for warehousing purposes, both in terms of storage capacity and response time.

In the remainder of this section, we present the global architecture of our medical data warehouse, two examples of simple datamarts (i.e., datamarts that store "simple" data), an example of complex datamart (i.e., storing complex data), as well as implementation issues.

Global Architecture

Our data warehouse is organized as a collection of interconnected datamarts sharing common dimensions. Each datamart stores the data related to a given medical field (e.g., biological analysis results, biometrical

Related Content