Data warehousing and data mining – an overview

Dr. Suman Bhusan Bhattacharyya

MBBS, ADHA, MBA

Abstract

With continuous advances in technology, increasing number of clinicians are using electronic medical records to accumulate substantial amounts of data about their patients with the associated clinical conditions and treatment details. The 'hidden' relationships and patterns within these information would further our medical knowledge including its efficiencies and deficiencies. Methodologies that are being used in parallel industries with increasing effectivity need to be modified and applied to discover this knowledge. This paper discusses, at a high level, the various methodologies that may be used, along with the elaboration of the various terminologies associated with data warehousing and knowledge discovery in databases (KDD).

Keywords

Electronic Medical Records (EMR), relational database management systems (RDBMS), rule-based alerts and warnings, clinical protocols, Clinical data warehousing (CDW), Structured Query Language (SQL), star schema, online transactional processing (OLTP), online analytical processing (OLAP), metadata, clinical decision support systems (CDSS), evidence based medicine, outcomes analysis.

Introduction

Electronic medical records are becoming more ubiquitous in day-to-day clinical practice. They capture clinical data, store in personal database as well as mirror it in local and regional databases. Data capture, storage, retrieval and display are all performed. They also allow display of alerts, warnings, guide a clinician through a clinical protocol by way of workflow, and online transactional processing where "intelligent" data display is made through running structured queries using SQL, etc.

Unfortunately, all this data residing in a RDBMS is good enough for basically the following purposes with respect to improving patient care.

1. Display against a period of time allowing for better visualization of the patient's clinical condition
2. Potentially life-saving alerts/warnings about a patient based on the clinical information collected about the patient

They are not able to support either evidence based medicine or outcomes analysis directly. To perform these tasks, it is necessary to have a data warehouse or at least an appropriate custom-built interface for the same. Although running specially designed queries may be able to accomplish this task, the pay off is that the user needs to correctly design them and retrieving results proves to a slow process as the data is usually not "analysis-ready".

It does however hold the potential to unleash a revolutionizing wealth of information regarding disease processes, disease progression, best method of

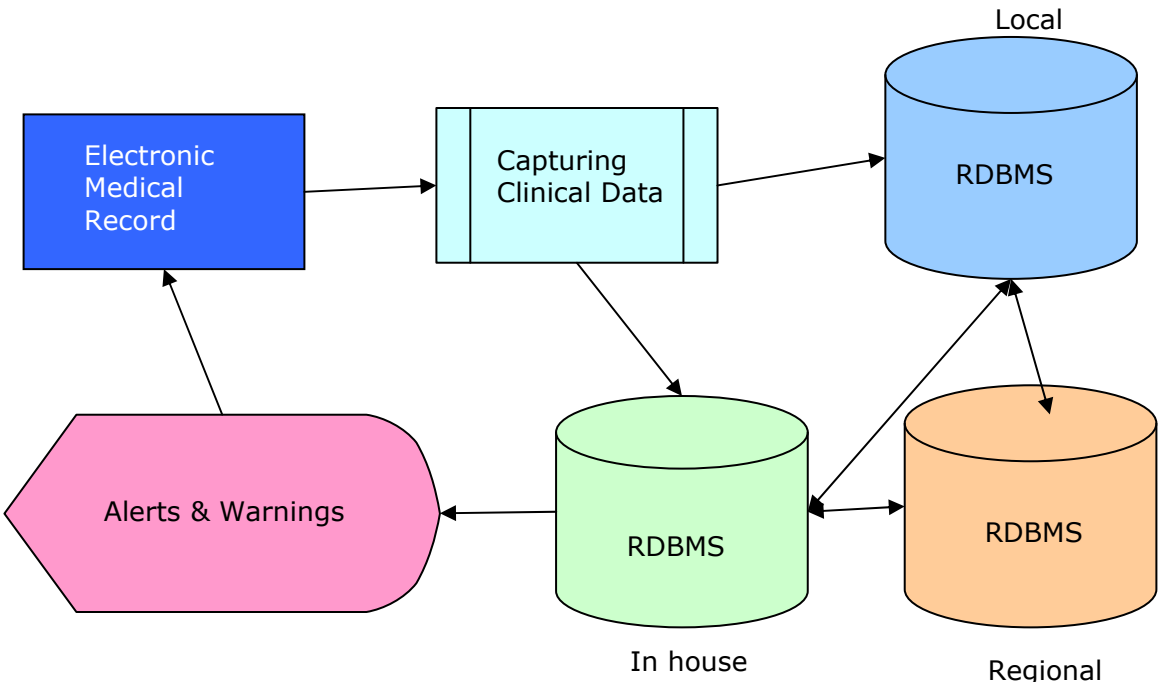
treatment, optimizing costs while maximizing efficiency, etc. Currently, the way to make this possible is to use online analytical processing by way of using data warehousing and data mining.

Data mining, functionally, is the process of discovering interesting knowledge from large amounts of data stored in various data repositories like databases or data warehouses. The process involves integration of techniques likes database technology, statistics, artificial intelligence, high performance computing, data visualization, image/signal processing, and spatial data analysis. By performing this process, interesting knowledge, patterns and high-level information can be extracted, viewed and browsed from multiple angles. The knowledge so discovered can be applied to decision-making, process control, information management, and query processing. [1]

Current Status

Electronic medical records capable of capturing clinical data, storing them locally (i.e. the clinician's own database), in-house (i.e., in the same organization like clinic, department or hospital), and regionally (i.e. in the same geographical area), capable of displaying data on request, alerts that are rule-based and patient-specific (display an alert if this patient's systolic blood pressure comes down below 60 mmHg or fasting blood sugar is more than 110 mg/dL on three consecutive days, etc.), warnings that have been pre-set (display a warning if any patient allergic to penicillin group of drugs and is suffering from rheumatic fever with ASO titer more than 200 Todd units, or a contra-indicated drug is prescribed, or two interacting drugs are prescribed concomitantly, etc.), following clinical protocols and performing other OLTP functions.

The relevant software architecture is as follows.



Displaying data
Displaying rule-based patient-specific alerts
Displaying pre-set warnings
Following clinical protocols
Online Transactional Processing (OLTP)

The Requirement

1. Discovering disease trends – Evaluation of stored data can lead to discovery of trends and patterns that would enhance the understanding of disease progression and management
2. Insurance risk assessment – In the future the insurance companies will clinically assess a person for the most likely risks for a specified period and then calculate the premium for health insurance. To do be able to do this, they will need to analyze the population health risks and then match them with the findings of the person being assessed. Such analyses may realistically only be done using clinical data warehouses

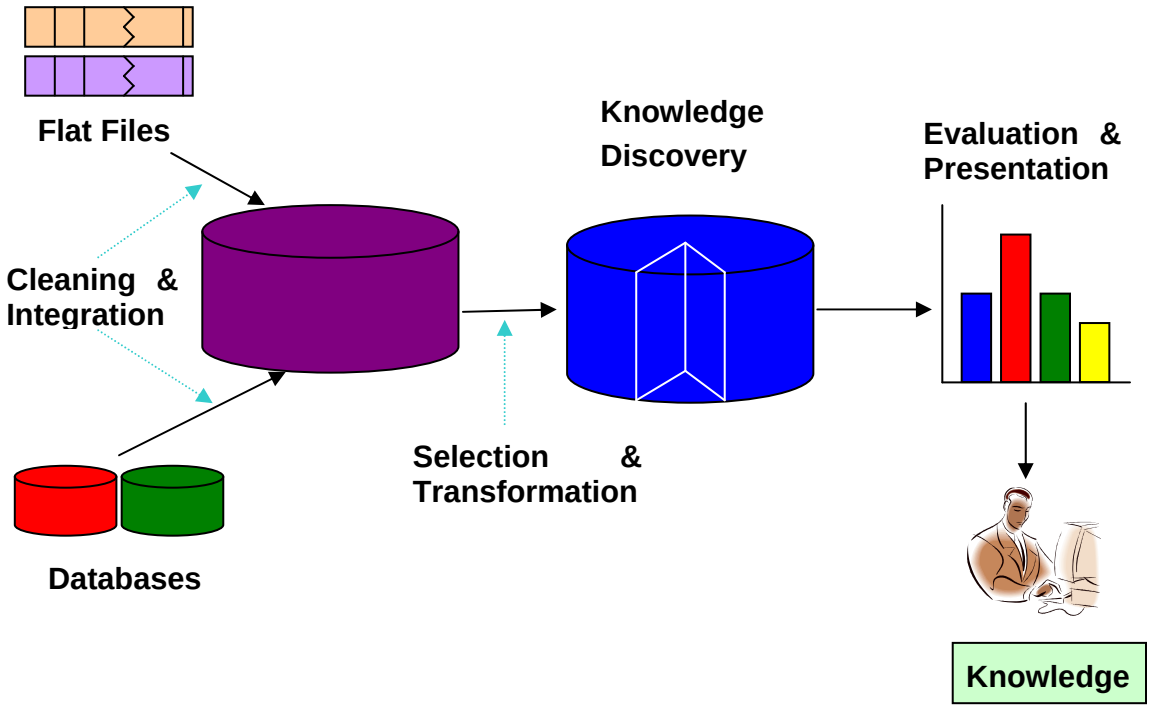
Methodology

Steps of Knowledge Discovery

1. Cleaning to remove data inconsistencies and aberrations (called "noise")
2. Extraction of data from multiple sources and integrating them into a data warehouse (it is a common practice to perform data cleaning and data integration as a pre-processing step before storing the resultant data)

3. Selection of data relevant to the analysis task by retrieving them from the data base
4. Transformation of data that are consolidated into forms appropriate for mining by performing summary or aggregation operations (occasionally data transformation and consolidation are performed before the data selection process, particularly in data warehouses)
5. Data mining is the essential process where intelligent methods are applied in order to extract data patterns
6. Evaluation of extracted data pattern is performed to identify the truly interesting patterns representing knowledge
7. Knowledge presentation is carried out by using visualization and knowledge representation techniques that are used to present the mined knowledge to the user

The process is pictorially depicted below. [1]



The process of getting data out of databases and into data warehouses is no easy task. This is the first step and the most time-consuming one. The next is to create, where necessary, data marts. This needs to be followed by data mining through framing appropriate queries and running them. The results display assumes vital importance for clinicians since only when a particular data is presented in a particular way does it become meaningful – a series of values is less preferable to a graphical display, while the value of density of a tissue with the image would convey more meaning than the image alone.

All the data clinical captured through electronic medical records can provide invaluable insights into the trends, progression, patterns and management of disease and its processes through the process of knowledge discovery using data analysis techniques.

The first step in the knowledge discovery process is to define the problem, followed by data modeling. Data models capable of addressing the problem needs to be built deployed and evaluated before actual use. This takes 50 – 90% of total project time. [3]

After the problem has been defined, the data is prepared using the data warehousing technique called ETL (extract, transform, and load). Extraction is typically gathering the data from multiple, heterogeneous and external sources. Transforming is converting the data from legacy or host RDBMS format to data warehouse format. Loading is sorting, summarizing, consolidating, computing, viewing, checking integrity, and building indices and partitions of the data.

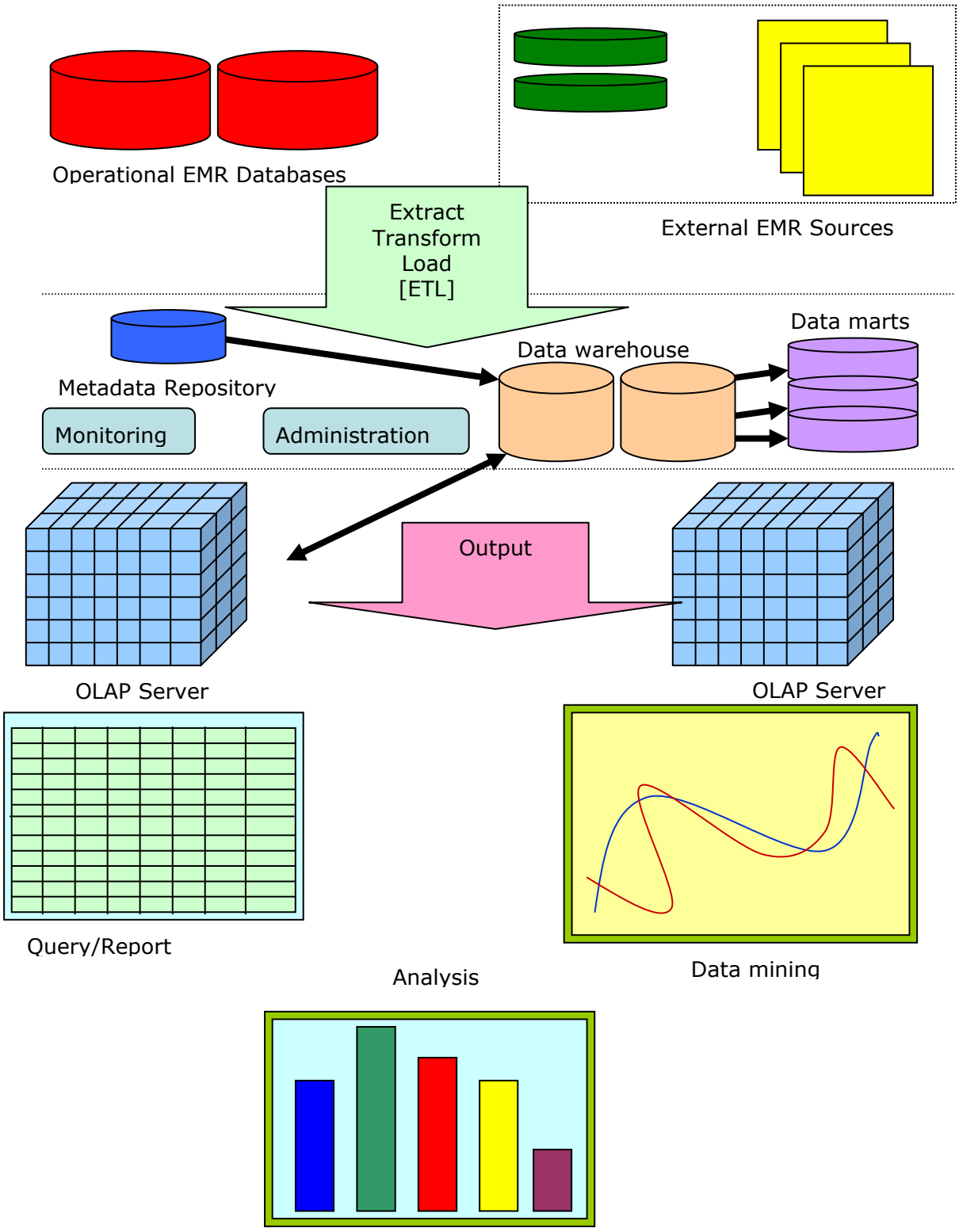
A data schema appropriate for multidimensional databases (like star or snowflake, etc.) is used to ultimately place the data in the clinical data warehouse. A metadata repository is created to increase the efficiency of the data mining process.

The entity-relationship data model is commonly used in the design of relational databases (RDBMS), where a database schema consist of a set of entities and the relationships between them. This is usually a two dimensional in nature. Such a data model is appropriate for OLTP. On the other hand, a data warehouse requires a concise, subject-oriented schema that facilitates OLAP. The most popular data model for a data warehouse is a multidimensional model (three dimensional or cubic, where the each data item has three or more contexts associated with it). Just as relational query languages like SQL is used to specify relational queries, a data mining query language (DMQL) is used to specify data mining tasks.

Metadata is data about data and they define data warehouse objects like data names, definitions, date created and captured for date time stamping any extracted data, etc. They are used as a directory to locate contents, for data mapping, and a guide to the algorithms for various types of summarized data. These need to be stored and managed persistently. [1]

Once all these have been successfully completed, the data is analyzed to discover knowledge. These are presented as reports, graphical analysis and pictorial depiction of patterns or trends. Such information so presented is then studied by the experts to gain knowledge and, where necessary, run further queries to drill down or up the data unmasking further knowledge. Careful harnessing of this technique by qualified analyzers can and do result in the discovery of hitherto unknown patterns leading to a wealth of new information that radically changes thinking of a number of axioms accepted as sacred facts. This has been proven to be the case in a number of business areas and health care as well. Not only are business processes rejuvenated but clinical processes that are strikingly similar to business process at a conceptual level, are too.

The process is pictorially depicted below. [1]



Pressures of Future

Conceptually, electronic medical records are already passé. The next generation is clinical knowledge discovery that lie hidden within them through clinical data warehousing. Mother Nature has forever thrown challenges to us humans that

need to be overcome in order to progress on to the next level of human development. Clinical data is already available and where there is time and properly trained computer savvy epidemiologists and bio-statisticians available, this should not be a problem *per se*. But, proper data collection, authentication, validation, and analysis of the data are a Herculean task in itself. Things can certainly be made more streamlined and less time consuming through automating these tasks. There is a further point to be considered. Many a life is lost due to unavailability of life-saving information at the right time. Systems that can provide vital knowledge regarding the most likely diagnosis or the "right" treatment would be a most welcome addition to any clinician's repertoire. Current clinical decision support systems rely on clinical data warehouses in some form or the other where mostly static data is processed and stored beforehand for ready analysis as and when required. Dynamic data that is processed and analyzed in real time is going to be the next "in thing".

Current Practices

Currently the following applications are being used in clinical settings for clinical data warehousing. Suffice it to say, learning to use them requires considerable effort and understanding. Custom-made solutions hardly exist.

1. Cognos
2. Business Objects
3. Teradata
4. Informatica
5. Oracle Discoverer
6. SPSS
7. SAS Tools
8. Epi Info with Epi Report
9. Darwin
10. Clementine [2]

Conclusion

The concept, usefulness and practicality of data warehousing is already a fact. They are being tested and rolled out in increasing numbers world wide. [2] The return-on-investments have well justified the costs involved in building and maintaining one. It is but a natural progression on the path of information and knowledge management.

In a clinical setting, the first step is to capture, authenticate, validate, store and retrieve the data in the proper manner. This is already being done through electronic medical records. The next is to unearth the knowledge that lies "hidden" within the captured data.

This is accomplished through clinical data warehousing and data mining using a team of health analysts. This team is made up of medical specialists and data mining technology experts. While the specialists help by framing the "right" questions for analysis, the technologists do the actual data model designing and the tasks of ETL and data mining. The results are passed back to the specialists who then perform the task of discovery and report the findings to the concerned persons.

References

1. Jiawei Han, Micheline Kamber. *Data mining – concepts and techniques*. Morgan Kaufmann Publishers. ISBN – 1055860-489-8
2. Philip Baylis et al. *Better health care with data mining, Clementine – working with health care*. SPSS white paper. Shared Medical Systems Limited, UK
3. Kristin B. Degrug, MSHS. *Healthcare Applications of Knowledge Discovery in Databases*. Journal of Healthcare Information Management, Vol. 14, no. 2, Summer 2000
4. Dale Sanders. *Healthcare Analytics: Standing on the Brink of a Revolution*. Journal of Healthcare Information Management, Vol. 16, no. 4
5. Paul Kallukaran, Jerry Kagan. *Datamining at IMS HEALTH How we turned a mountain of data into a few information-rich molehills*. IMS HEALTH, Plymouth Meeting, PA USA
6. Jonathan C. Prather, David F. Lobach, et al. Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse
7. Craig S. Ledbetter, Matthew W. Morgan. Toward Best Practice: Leveraging the Electronic Patient Record as a Clinical Data Warehouse
8. Micheal Silver, Taiki Sakata, et al. Case Study: How to Apply Data Mining Techniques in a Healthcare Data Warehouse. Journal of Healthcare Information Management, Vol. 15, no. 2, Summer 2001