Data Mining on Distributed Medical Databases: Recent Trends and Future Directions

Yasemin Atilgan and Firat Dogan

Dogus University, Computer Engineering Department, Research Assistant
Acibadem, Istanbul, Turkey
{yatilgan,fdogan}@dogus.edu.tr

Abstract. As computerization in healthcare services increase, the amount of available digital data is growing at an unprecedented rate and as a result healthcare organizations are much more able to store data than to extract knowledge from it. Today the major challenge is to transform these data into useful information and knowledge. It is important for healthcare organizations to use stored data to improve quality while reducing cost. This paper first investigates the data mining applications on centralized medical databases, and how they are used for diagnostic and population health, then introduces distributed databases. The integration needs and issues of distributed medical databases are described. Finally the paper focuses on data mining studies on distributed medical databases.

Keywords: Data mining; Distributed Databases; Medical Databases.

1 Introduction

Today like all organizations healthcare organizations are also employing some sort of information systems. These systems provide technology to store data which take the form of numbers, text, chart or images. As a result large medical databases have grown in today's healthcare environment. These large medical databases store vast amount of data that is not used. It is an important issue in Information Age to extract knowledge from these data repositories, especially while healthcare organizations are facing a major challenge on improving service of quality delivered at affordable costs. The strategy to lower cost, raise quality and still be competitive in the information age is to build up strong healthcare information systems for knowledge management and decision support. Knowledge is the new capital of organizations. Today the economy is knowledge economy.

Using computer and information technologies in healthcare services helps to achieve efficiency, effectiveness in diagnostic decision making, cost economy, better risk management and strategic planning in a competitive healthcare environment [1].

Organizations are much more able to store data than to extract knowledge from it. Data modeling and analysis tools like data mining are widely used to generate knowledge rich environment. Using data mining techniques on medical data helps better decision making in diagnosis, better care on patients, finding way of preventing

M. Ulieru, P. Palensky, and R. Doursat (Eds.): IT Revolutions 2008, LNICST 11, pp. 216–224, 2009.

some diseases and early prevention in some medical areas, and all these provide cost reduction, better risk management and quality of services in healthcare services.

The improvements in information technologies have not only brought services to store data. Before the revolution in computer systems, most organizations had only a handful of computers, and for lack of a way to connect them, these operated independently from one another. The advances in microprocessors and networks changed that situation. The result of these technologies brought distributed systems in contrast to previous centralized systems [2]. Tanenbaum and Steen define distributed system as a collection of independent computers that appear to the users of the system as a single computer. Distributed systems made the connection of databases that are geographically or physically separate from each other possible. The developments in distributed technology also made information sharing possible. A shared information system can be considered as a series of computer systems - each is also a data repository - interconnected by some sort of communications network. Figure 1 visualizes a centralized database system and distributed database system.

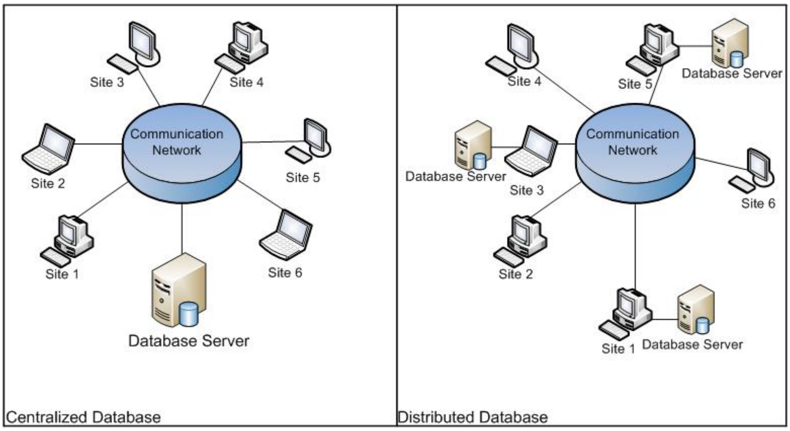

Fig. 1. Schematic representation of centralized and distributed databases

As a result of above mentioned developments in technology brings the concept of distributed data mining. The medical data is distributed because the data repositories are located in different hospital or different departments in one hospital and heterogeneous because patient’s medical records can be stored in databases in different forms, inspecting results can be saved as imaging files, clinical cases or doctor’s advices can be saved as text documents, and other video or images. Today the challenge is to extract knowledge from these distributed and heterogeneous data repositories. Using data mining techniques on data stored in centralized databases is common. It is possible to find applications of different data mining techniques on different kind of medical data, which is naturally stored in centralized databases or the data used has been centralized to accomplish the study. The application of data mining on distributed databases is still not a common practice.

The purpose of this paper is to review recent data mining trends on distributed databases, give an insight about the studies and applications, and finally explore the future directions in distributed medical databases.

The remainder of the paper is organized as follows. Section 2 provides a brief overview of data mining perspective on medical databases and data mining applications in this area; section 3 explores the researches and studies about data mining on distributed medical databases.

2 Data Mining on Medical Databases

Data mining is a field that uses database technology, statistics, pattern recognition, data visualization, machine learning and expert systems for knowledge discovery. A database is a collection of data that is organized so that its contents can easily be accessed, managed and updated [3]. In today's healthcare environment vast amount of clinical and personal data about patients are kept in medical databases which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of information hidden in these data that is largely untapped [4]. The challenge is to transform these data into information and knowledge. Data mining tools make knowledge discovery possible from these large data repositories.

Frawley et al. define knowledge discovery as the nontrivial extraction of implicit, previously unknown, and potentially useful information from data:

> *"Given a set of facts (data) F, a language L, and some measure of certainty C, we define a pattern as a statement S in L that describes relationships among a subset FS of F with a certainty c, such that S is simpler (in some sense) than the enumeration of all facts in FS. A pattern that is interesting (according to a user-imposed interest measure) and certain enough (again according to the user's criteria) is called knowledge. The output of a program that monitors the set of facts in a database and produces patterns in this sense is discovered knowledge."*

A lot of valuable knowledge hidden in the databases can be discovered using data mining approach, which is worth exploring. For example, to understand the relationships between characteristics of patient symptoms and the illness so that patients can utilize the results of this research to assist in guiding patients to connect their own symptoms to the type of illness accurately large data sets and classification methods in data mining technique can be used [6]. Applications of data mining techniques in health care systems vary from heart disease prediction [4] to early detection [7] or risk grouping [8] of prostate cancer, from predicting breast cancer survivability [9], to assessing healthcare resource utilization of lung cancer patients [10]. Kraft et al. studied for a process of knowledge generation for predicting the length of stay of spinal cord injury patients. They used nursing diagnosis to predict length of stay. As a result no one diagnostic cluster was found as the critical factor in the prediction model. They claim that the decision of data to be warehoused becomes very important for better application of data mining techniques and improve knowledge discovery. This also shows that data warehousing and the design issues in databases are also important, but this is a subject of another study. Chae et al.

presented analysis of healthcare quality indicators using data mining techniques. They identified the important factors influencing the inpatient mortality. They demonstrated how a data mining technique could be used in developing continuous quality improvement strategy. In literature it is possible to find other studies and applications implementing different data mining techniques using data from different areas in healthcare [13][14][15][16]. Some studies show that by accessing the right information, at right time at right place enables taking the right action on time or helps early prevention, while some studies emphasis the importance of quality and availability of data. Accessing the right knowledge provides diagnosing and treating patients, as well as to prevent and maintain some illnesses [17].

3 Distributed Medical Databases

3.1 Integration of Medical Databases

As healthcare institutions are hiring new information systems the data being digitalized is increasing exponentially. These information systems are classified according to place they are used and the type of data stored. Generally they are classified as Hospital Information Systems (HIS), Radiology Information Systems (RIS), Laboratory Information Systems (LIS), Picture Archiving and Communication System (PACS) and so on[18]. All these systems are centralized stand-alone systems, so they work independently from each other. As a result by the nature of the data type stored in these systems heterogeneity is common. The other issue in medical databases is, since the healthcare organizations are located in different places and have their own data repositories the medical data are stored in distributed databases by nature. So, distributed medical databases in this study refer to the medical databases that are distributed because of their nature. Designing a distributed database for a geographically dispersed organization is another issue.

Today, the healthcare systems around the world are moving towards the integration of these distributed data sources, to improve the efficiency in healthcare services through data sharing [19]. Building information systems without carefully planning and developing independent systems for new services, with medical data of a patient scattered in various stand-alone databases, causes cumulative inefficiencies in the information processing and sharing, thus leading to medical mishaps and potential risks in legal liabilities from patient care delivery [20]. As a result, integration of existing heterogeneous databases to provide effective and efficient knowledge discovery and better knowledge management in hospitals is one of the most popular issues in healthcare informatics.

Although database integration issues are not new to database researchers, hospitals have many unique characteristics which entail special integration design considerations for the following reasons [1]:

- Operational efficiency
- Cost economy
- Effective diagnostic decision making
- Emergency care services
- Legal requirements

- Strategic planning and risk management
- Education and research

They proposed a virtual database integration approach for the database integration needed in a hospital to solve data redundancy, data inconsistency, data model incompatibility, time delay and inconvenience problems and satisfy the information handling needs from various hospital services and functions and also retain the autonomy of departmental system development and provide central planning and control mechanism for system expansion. Integration is also important for multi-site healthcare organizations. These organizations implement heterogeneous information management systems interacting with distributed databases [21]. The study explores the ‘middleware services’ as a solution to guarantee data exchange across different types of applications and database management systems, and reduce the costs of systems development and modification. Their study concluded that their design architecture is a highly effective tool for developing reusable object/components that can substantially reduce the time and effort required for system development and maintenance. In [22] the integration problem of heterogeneous data in medicine is studied. They present a concept and solution to support intra and also inter-institutional integration of systems that are being used in healthcare institutions. Their concept was based on requirements of researchers and clinicians to integrate data from various existing component systems. They represented a “light-weight” approach to interconnect existent data sources, because of legal and regulatory restrictions. Another study about integration is given in [23]. They developed various methods and tools for database integration, because they mention that different biological and clinical research projects are based on collaborative efforts among international organizations. The method they present detects the inconsistencies automatically and stores the corresponding transformations in a formal structure to support knowledge discovery professionals. In literature it is possible to find different kinds of integration solutions like a framework used to integrate molecular biology database using context graph [24], an architecture based on multi-agent system and grid technology [25] and other studies for the integration of genomic data [26][27].

Integration of medical databases is an important issue and necessary for knowledge discovery in databases. Even it is not discussed in this study, also legal and privacy issues are important and should be considered during integration process.

3.2 Data Mining Applications on Distributed Medical Databases

Since modern medicine generates almost daily huge amounts of heterogeneous data and the stored medical data may contain images, signals, clinical information, cholesterol levels, etc., as well as the physician's interpretation [28], researchers present architecture for cooperative work in heterogeneous medical information, such as Hospital Information System(HIS) and Laboratory Information System (LIS) [29].

A methodology and operational framework for applying data mining techniques from distributed and heterogeneous clinical data sources is presented by [30]. In their study they state that epidemiological studies are important for health prevention and health prevention is highly dependent on information transfer. To solve the problem of mining heterogeneous and distributed data sources they indicate that a multi-phase data integration procedure should be followed. After the integration methodology, they propose their study on the specifics of knowledge discovery processes. They particularly study on discovery of interesting associations between the recorded patients' clinical data items. Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data [31]. They make the assumption that these associations may be linked with indicative epidemiological and health-indicators. Their results show the effectiveness of respective distributed data mining operations. As a result they discover and form interesting associations according to specific query posted by the user via the patient clinical data directory service. Another study presents an intelligent agent based framework for knowledge discovery comprising multiple heterogeneous healthcare data resources [32]. They argue that with the existence of multiple heterogeneous data repositories in a healthcare enterprise a distributed data community, such that any data mining effort draws upon the 'holistic' data available within the entire healthcare enterprise should be established. The proposed multi Agent-Based Data Mining Info-Structure (ADMI), uses the advantage of a multi-agent architecture which features the amalgamation of various types of intelligent agents, each responsible for an independent task. It is responsible for the generation of data-mediated diagnostic-support and strategic services. They designed the Interface Agent (IA) to collect user-specification for a data mining service via a web-based interface, Data Collection Agent (DCA) to facilitate the on-demand retrieval of relevant data from the multiple healthcare data repositories, Data Mining Agent (DMA) for coordinating the entire data mining activities, and Services Generation Agent (SGA) to process the data mining results produced by DMA to generate decision-support or strategic services as per the user's request. They demonstrate that autonomous, reactive and proactive intelligent agents provide an opportunity to generate end-user oriented, packaged, value-added decision-support/strategic planning services for healthcare professionals and managers. Their work was leverage and a prototype version of their framework is under development.

As the volume of data stored increase, new challenges for their effective understanding raise. Since the processes and activities that are computationally intensive, collaborative and distributed in nature are involved in knowledge discovery in large data repositories, high level frame works like Knowledge Grid are developed [33]. Grid computing aims to aggregate distributed computing resources, hide their specifications and present a homogenous interface to end users for high performance or high throughput computation [34]. The grid can play a significant role in providing an effective computational infrastructure support for data mining from very large datasets maintained over geographically distributed sites by using computational power of distributed systems [35].

Grid technology can be used in hospitals and medical related works to share and integrate heterogeneous medical sources [36]. The study also introduces the related works on medical grid applications. The applications introduced are mostly related with databases storing medical images like CT, MRI and mammograms. Since the digital medical images represent a tremendous amount of data, a hospital is capable of producing several Terabytes of medical image data each year. Because of this reason using grid technologies for medical image databases is more common.

The DataMiningGrid system is a system that has been designed and developed in order to meet the requirements of modern and distributed data mining scenarios. The system was recently built on top of existing Globus technology inter alia to address the requirements of a community of medical users and enable them to perform on-the-fly analysis of geographically distributed medical databases. DataMiningGrid system provides tools and services facilitating the grid-enabling of various applications including data mining and statistical applications without major intervention on the application side [37]. The details of DataMiningGrid system architecture is introduced with details in [38]. The study also presents the grid-enabling data mining applications and the requirements. It is claimed that there is still a long way to go before grids can be widely used in the medical domain, however prototypes and experiments are developed on medical applications [39].

4 Conclusion

The studies show that there is variety of data mining applications on medical database. These studies have been conducted on different subjects in healthcare like cancer prediction, length of hospital stay optimization and so on. In literature different data mining techniques applied to centralized information repositories, and some studies make the comparison between these techniques. We can say that there is a lack of data mining applications on distributed medical databases. The recent trends on data mining on distributed medical databases are mostly in design base, or applied in sample databases. The future directions on database and data mining applications is designing more scalable and internet based data mining architectures so that the data will be available anytime by everyone who has access to Internet.

References

1. Liu Sheng, O.R., Garcia, H.-M.C.: Information Management in Hospitals: An Integrating Approach. In: 9th IEEE International Phoenix Conference on Computers and Communications, pp. 296–303. IEEE press, Scottsdale (1990)
2. Tanenbaum, A.S., Steen, M.: Distributed Systems: Principles and Paradigms, pp. 2–3. Prentice Hall, New Jersey (2002)
3. Obenshain, M.K.: Application of Data Mining Techniques to Healthcare Data. Infection Control and Hospital Epidemiology, 690–695 (2004)
4. Palaniappan, S., Awang, R.: Intelligent Heart Disease Prediction System Using Data Mining Techniques. In: IEEE/ACS International Conference on Computer Systems and Applications, pp. 108–115. IEEE press, Doha (2008)
5. Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J.: Knowledge Discovery in Databases: An Ovrview. AI Magazine, 57–70 (1992)
6. Chang, C.L.: A Study of Applying Data Mining to Early Intervention for Developmentally-delayed Children. Expert Systems with Applications, 407–412 (2007)
7. Zhang, Z., Zhang, H.: Development of a Neural Network Derived Index for Early Detection of Prostate Cancer. IEEE International Joint Conference on Neural Networks, 3636–3641 (1999)
8. Churilv, L., Bagirov, A.M., Schwartz, D., Smith, K., Dally, M.: Improving Risk Grouping Rules for Prostate Cancer Patients with Optimization. In: IEEE Proceedings of the International Conference on System Sciences, Hawai (2004)

9. Delen, D., Walker, G., Kadam, A.: Predicting Breast Cancer Survivability: a comparison of three data mining methods. Artifical Intelligence in Medicine, 113–127 (2005)
10. Phillips-Wren, G., Sharkey, P., Dy, S.M.: Mining Lung Cancer Data to Assess Healthcare Resource Utilization. Expert Systems with Applications, 1611–1619 (2008)
11. Kraft, M.R., Desouza, K.C., Anndrowich, I.: Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population. In: IEEE Proceedings of the International Conference on System Sciences, Hawaii (2002)
12. Chae, Y.M., Kim, H.S., Tark, K.C., Park, H.J., Ho, S.H.: Analysis of Healthcare Quality Indicator Using data Mining and Dceision Support System. Expert Systems with Applications, 167–172 (2003)
13. Lee, S., Abbott, P.A.: Bayesian Networks for Knowledge Discovery in Large Datasets: Basics for Nurse Researchers. J. Biomedical Informatics, 389–399 (2003)
14. Lin, F., Chou, S., Pan, S., Chen, Y.: Mining Time Dependency Patterns in Clinical Pathways. In: IEEE Proceedings of the International Conference on System Sciences, Hawaii (2000)
15. Wilson, A.M., Thabane, L., Holbrook, A.: Application of Data Mining Techniques in Pharmacovigilance. British Journal of Clinical Pharmacology, 127–134 (2003)
16. Silva, A., Cortez, P., Santos, M.F., Gomes, L., Neves, J.: Mortality Assessment in Intensive Care Units via Adverse events Using Artificial Neural Networks. Artificial Intelligence in Medicine, 223–234 (2006)
17. Goodwin, L., VanDyne, M., Lin, S., Talbert, S.: Data Mining Issues and Opportunities for Building Nursing Knowledge. J. Biomedical Informatics, 379–388 (2003)
18. Ahn, C., Nah, Y., Park, S., Kim, J.: An integrated medical information system using XML. In: Kim, W., Ling, T.-W., Lee, Y.-J., Park, S.-S. (eds.) Human Society Internet 2001, vol. 2105, pp. 307–322. Springer, Heidelberg (2001)
19. Au, R., Croll, P.: Consumer-Centric and Privacy-preserving Identity Management for Distributed e-Health Systems. In: IEEE Proceedings of the International Conference on System Sciences, pp. 1–10 (2008)
20. Troyer, G.T., Salman, S.L.: Handbook of Health Care Risk Management. Aspen Systems Corporation, Maryland (1986)
21. Chu, S., Cesnik, B.: A three-tire Clinical Information Systems Design Model. International J. Medical Informatics, 91–107 (2000)
22. Wurst, S.H.R., Lamla, G., Schlundt, J., Karlsen, R., Kuhn, K.A.: A Service-oriented Architectural Framework for the Integration of Information Systems in Clinical Research. In: IEEE Proceedings of the International Symposium on Computer-Based Medical Systems, pp. 16–163 (2008)
23. Anguita, A., Perez-Ray, D., Crespo, J., Mojo, V.: Automatic Generation of Integration and Preprocessing Ontologies for Biomedical Sources in a Distributed Scenario. In: IEEE Proceedings of the International Symposium on Computer-Based Medical Systems, pp. 336–341 (2008)
24. Khan, N., Rahman, S., Stockman, A.G.: A Framework for Molecular Biology Database Integration Using Context Graph. In: IEEE Proceedings of the International Symposium on Computer-Based Medical Systems, pp. 21–26 (2004)
25. Di Lecce, V., Amato, A., Calabrese, M.: Data Integration in Distributed Medical Information Systems. In: Canadian Conference on Electrical and Computer Engineering, pp. 1497–1502 (2008)
26. Gros, P.E., Herisson, J., Ferey, N., Gherbi, R.: Combining Applications and Databases Integration Approaches in a Common Distributed Genomic Platform. In: IEEE Proceedings of the International Conference on Advanced Information Networking and Applications, pp. 433–438 (2005)

27. Douthart, R.J., Pelkey, J.E., Thomas, G.S.: Database Integration and Visualization of Maps of the Human Genome Using the GnomeView Interface. In: IEEE Proceedings of the International Conference on System Sciences, pp. 49–57 (1994)
28. Cios, K.J.: Medical Data Mining and Knowledge Discovery. Studies in Fuzziness and Soft Computing. Physica - Verlag (2001)
29. Li, K., Yao, D.: Cooperative Work in Heterogeneous Medical Information Systems. In: IEEE Proceedings of the International Conference on Communications, Circuits and Systems (2006)
30. Potamias, G.A., Moustakis, V.S.: Knowledge Discovery from Distributed Clinical Data Sources: The Era for Internet-based Epidemiology. In: IEEE Proceedings of EMBS International Conference, pp. 3638–3641 (2001)
31. Han, J., Kamber, M.: Data Mining Concepts and Techniques. Morgan Kaufmann, US (2001)
32. Zaidi, S.Z.H., Abidi, S.S.R., Manickam, S.: Distributed Data Mining from Heterogeneous Healthcare Data Repositories: Towards an Intelligent Agent-Based Framework. In: IEEE Proceedings of Symposium on Computer-Based Medical Systems (2002)
33. Congiusta, A., Talia, D., Trunfilo, P.: Distributed Data Mining Sevices Leveraging WSRF. Future Generation Computer Systems, 34–41 (2007)
34. Luo, P., Lü, K., Shi, Z., He, Q.: Distributed Data Mining in Grid Computing Environments. Future Genertion Computer Systems, 84–91 (2007)
35. Luo, J., Wangc, M., Hud, J., Shia, Z.: Distributed Data Mining on Agent Grid: Issues, platform and development kit. Future Generation Computer Systems 3, 61–68 (2007)
36. Zheng, R., Jin, H., Zhang, Q., Liu, Y., Chu, P.: Heterogeneous Medical Data Share and Integration on Grid. In: IEEE Proceedings of the International Conference on BioMedical Engineering and Informatics, pp. 905–909 (2008)
37. Jarm, T., Kramar, P., Županič, A. (eds.) : Medicon 2007. IFMBE Proceedings 16, 166–169 (2007)
38. Stankovski, V., Swain, M., Kravtsov, V., Niessesn, T., Wegener, D., Kindermann, J., Dubitzky, W.: Grid-enabled Data Mining Applications with DataMiningGrid: An architectural perspective. Future Generation Computer Systems, 1–21 (2007)
39. Montagnat, J., Breton, V., Magnin, I.E.: Using grid Technologies to Face Medical Image Analysis Challenges. In: IEEE/ACM 3rd International Symposium on Cluster Computing and the Grid, pp. 1–5 (2003)