International Journal of Bio-Science and Bio-Technology
Vol.7, No.3 (2015), pp.217-222
http://dx.doi.org/10.14257/ijbsbt.2015.7.3.23

# A Framework for Data Warehouse Using Data Mining and Knowledge Discovery for a Network of Hospitals in Pakistan

Muhammad Arif [1,2], Asad Khatak[2] and Mehdi Hussain[1,3,]

*[1]Faculty of Computer Science and Information Technology, University of Malaya 50603 Kuala Lumpur, Malaysia*
*[2]Computer Science Department, Comsats Institute of Information and Technology Islamabad Pakistan*
*[3]School of Electrical Engineering and Computer Science, National University of Sciences and Technology, Islamabad Pakistan*

## Abstract

*Now-a-days, in Pakistan especially in strategic, military and private sector hospitals, there is an increasing use of hospital information systems. It has been seemed that a localize approach for developing HMIS is prevailing i.e. conventional use of relational database system, which are isolated from others hospitals or remote data collection centers. In this, paper a framework is proposed for establishment of data warehouse to centralize data from remote hospitals and collection centers. Data mining and knowledge discovery modal is also discussed.*

**Keywords:** *Data warehouse, Data Mining, Knowledge Discovery (KDD), Hospital Information System, Online Transaction Processing (OLTP), Online Analytical Processing (OLAP).*

## 1. Introduction

One of the modern basic needs of a human being is quality health care system since it improves quality of life. That's why, public and private sector hospitals need to provide and improve their facilities and services. But, there seems a great deficiency in public sector hospitals in Pakistan; however, strategic, military and private hospitals have shown great improvement. One of the factors in their improved healthcare system [4,14] is the use of information technology [6, 7, 8,14]. IT has major positive impact in almost every part of life, therefore, increasing use of hospital information systems [1,3,8,9,14] in Pakistan is way forward to better future sine it will provide a timely and accurate information regarding patients care.

Recent episodes of dengue favor, polio and hepatitis has produced alarming situation in Pakistan and now it has become a main focus of public and private sector domain experts to predict them in specific regions. Also, there is increasing awareness of using statistical approaches to analyze these diseases. The prediction requires modern data mining techniques [1] having a reliable data center for better decision making. Unfortunately, public sector is not aware of the information system but strategic, military and private sector is making use of health care information system with limited use, based on localize system within premises [2,3,7,9,12]. There is need to shift from OLTP to OLAP i.e. use of data warehouse technology [2] to mitigate the situation. The possibility of centralize database is always considered but it seems impossible keeping in view the shortage of energy in Pakistan. Decentralize data centers may cater the situation better with an offline or online data updating approach. Also, complex statistical analysis on patient data require establishment of data warehouse [9, 10, 11,13].

In this paper, we will propose a frame work for data warehouse along with discussion on data mining and knowledge discovery techniques for a network of hospitals. Data

ISSN: 2233-7849 IJBSBT

mining [1] and statistics are co-related and specialized techniques could be employed for finding patterns and knowledge discovery within stored data. A network of hospitals will provide a lot of data to cater the data mining requirements.

## 2. Why Data Warehouse (DWH)

Establishment of DWH may have following advantages over OLTP relational databases:

1. Regardless to data store DWH provide common data model.
2. Inconsistence data is identified and resolved prior to uploading data to DWH.
3. DWH can store data for long period of time.
4. Traditional OLTP system cannot meet the complex data analysis techniques since data is organized in normal forms.
5. Data in DWH is presented without use of normalization, therefore, pattern recognition and knowledge discovery is possible with the use of complex statistical and data mining tools.
6. Decision making become easy with proper data reporting formats.
7. Multiple data sources can be effectively integrated by DWH.

### 2.1 Architecture of DWH

Different components integration of DWH can be seen in Fig.1.

The architecture of proposed framework will have three layers:

1. Data storage Layer which handles information storage.
2. Web and Application layer which handles uploading of data from multiple sources, give service to end user queries, security of data, also provide data analysis opportunity.
3. User Layer which handles the interface with the user

Fig.2 shows how these three layers will make a practical system. Different hospitals will have local data centers and they will communicate with each other via web service. This arrangement make it possible to use centralize database along with the integration of distributed medical resources.

### 2.2 Data Mining and Patient Record System (PRC)

In Pakistan, the use of computerizes patient record system is very new and in preliminary stage. Manual data analysis techniques are old and are not reliable too. Use of complex statistical analysis is very rare. Also, the health care system is in poor condition. However, the recent use of patient record system (PRC) in military, strategic and private hospitals may become a source for foundation of data mining and knowledge discovery.

DWH make it possible to gather huge information [6, 7, 8] at either centralize or distributed location. Further, processing of data will make it possible to extract meaningful data. Hospitals in Pakistan are taking care of huge amount of patients. A lot of data can be gathered for analysis and thus may become a major area for data mining. Meaningful patterns can be extracted by use of knowledge discovery techniques (KDD) and can be used in quality care of patients.

KDD stands for Knowledge Discovery in Databases and data mining is the back bone of KDD. Regularities and data patterns in raw data are discovered with KDD.

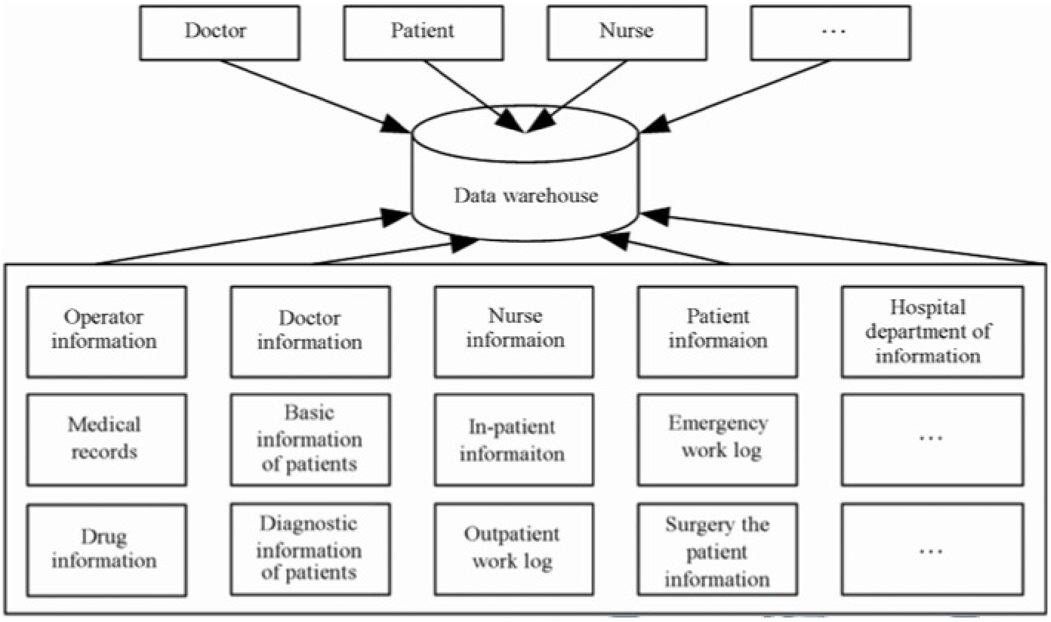

**Figure 1**

### 2.3 Principles of Data Mining

Following are the five steps for data mining:

1. Data Collection
2. Use of Data warehouse, databases and flat files in some cases
3. Use of statistical and mining Techniques
4. Finding patterns and clusters etc.
5. Application of Result

The word data mining clearly depicts the requirement of data to be gathered either through DHW or databases or spreadsheets. Since health care systems have the responsibility to handle patient data with secrecy, therefore, great care is required in gathering data. Of course, this cannot be possible without the help of doctors, nurses and hospital management.

Preprocessing of data is always required on raw data, so that, highest quality of data is assured. Errors and redundancies are removed. Patient identity and privacy is specially taken into account [6]. Different modeling techniques are available in data mining. Also, repeated permutations are used to get best data patterns. Data cleaning is job of great responsibility and cannot be accomplished without domain experts; however, 100% guaranteed results are not assured. Error removal can be both manual and automated and required joint effort of data mining experts, clinical practitioners and management e.g. out of range values can only be removed by clinical practitioners. High dimensionality of medical data is another important factor where clinical experts are required. It will not only reduce cost but also usefulness of patterns extracted. Supervised or unsupervised learning is an important factor. Data set selection and its attributes is very important and finally interpretation of results by domain experts is very crucial to success of knowledge discovery.

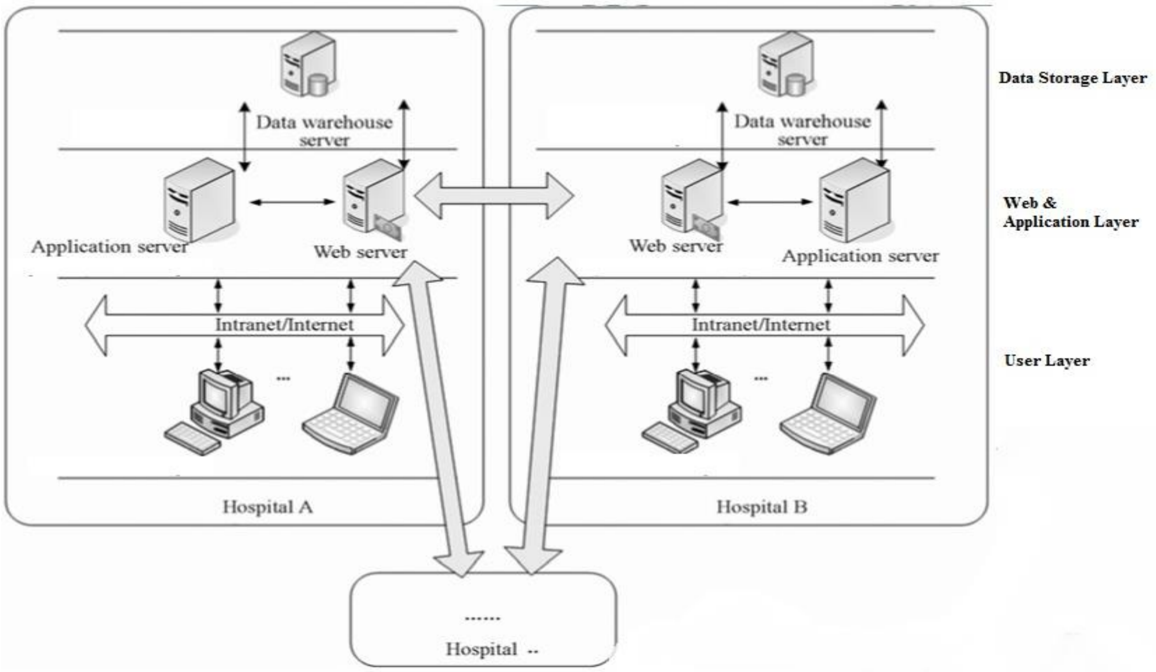

**Figure 2**

The major techniques used in data mining are Classification, Clustering and Association Rules. Supervised learning make use of classification e.g. categorization of medical images. Another example of supervised learning is pattern recognition based on an input pattern. Unsupervised learning makes use of clustering of data sets. It looks similar to classification but there is no predefinition of groups with data set. Uncovering of relationship between data is called association. It is a model that identifies specific types of data association e.g. retail sales community frequently use to identify items that are purchased together. Similarly physicians can obtain knowledge from huge data warehouse [2] with the help of information management experts.

### 2.4 Difference between Data Mining and Knowledge Discovery (KDD)

Data mining is extracting of patterns from data with the application of algorithms while knowledge discovery in databases uses additional step of evaluation and interpretation of patterns to make decision that qualify knowledge. KDD includes encoding schemes, preprocessing, sampling and projection of data.

### 2.5 Model of Knowledge Discovery in Databases (KDD)

Fig. 3 shows a KDD model and process. The KDD model consists of five stages.

1. Target Data Selection
2. Data Preprocessing
3. Data Transformation
4. Data Mining
5. Interpretation and Evaluation of Results

As mentioned earlier in DWH model, data for KDD may be made available from different resources which are the first step for KDD. Domain expert can use KDD tools to select initial set of medical data for analysis. Data preprocessing is required in KDD to cater for missing or incorrect data values. The initial data set may contain anomalous data involving different metrics and data types. Noisy data may lead to incorrect results and is necessary to be removed first. A policy decision is also required for missing values in data preprocessing. To get desired results, data miners use number of algorithms to transform the data into useful information. Interpretation of mined data results is the final and most important stage in KDD. This will lead to either accept or reject the results. In case of rejection, new attributes and instances are tried to get favorable results. Finally, the results are presented to the end user, so that, important decision can be made possible.

Since data is made available from different resources having different formats, therefore, data transformation to a common format is necessary for reliable interpretation of results. To get desired results, data mining use a

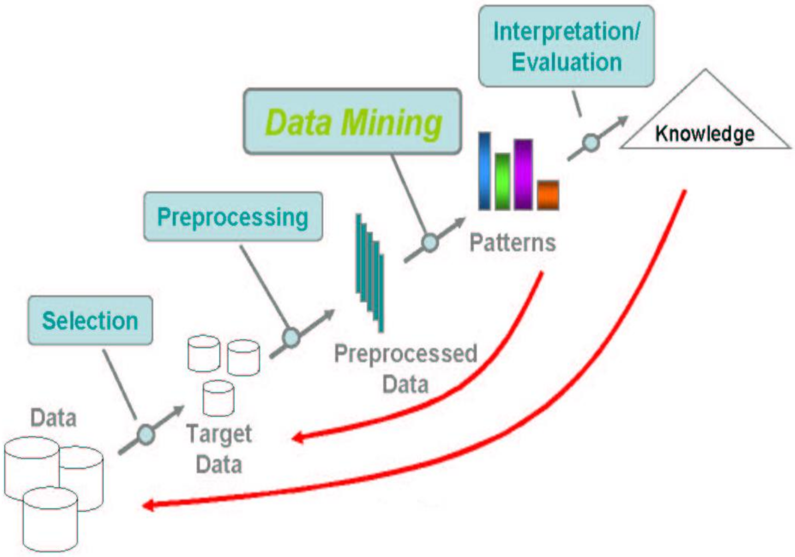

Figure 3

## 3. Conclusion

Establishment of data warehouse is need of day and a reliable solution for keeping huge amount of data for the purpose of analysis, storage, management, and information sharing and knowledge extraction. The sharing of data between different hospitals is the key concept and is viable through establishment of DWH at different location connected via web to a centralize location. DWH and knowledge discovery techniques may make it possible to take corrective measures, allow decision making, availability of latest information on disease control, facilitate diagnosis and treatment and finally promote medical services in an effective way to serve poor community of patients in Pakistan living in urban and remote areas.

## References

[1] S. Tsumoto and S. Hirano, "Data mining in hospital information system for hospital management. In Complex Medical Engineering", CME, ICME International Conference, IEEE, **(2009)**.

[2] M. F. Wisniewski, P. Kieszkowski, B. M. Zagorski, W. E. Trick, M. Sommers and R. A. Weinstein, "Chicago Antimicrobial Resistance Project", Development of a clinical data warehouse for hospital infection control, Journal of the American Medical Informatics Association, vol. 10, no. 5, **(2003)**, pp. 454-462.

[3] Y. H. Yan, L. J. Song, H. Xiong, H. Zhen, Z. T. Shu, C. Jie and Z. Tao, "Data mining analysis of inpatient fees in hospital information system", IT in Medicine & Education, ITIME, IEEE International Symposium, IEEE, **(2009)**.

[4] W. Wang, M. Wang and S. Zhu, “Healthcare information system integration: A service oriented approach”, Services Systems and Services Management, Proceedings of ICSSSM, International Conference, IEEE, **(2005)**.

[5] T. Xie, S. Thummalapenta, D. Lo and C. Liu, “Data mining for software engineering”, Computer, vol. 42, no. 8, **(2009)**, pp. 55-62.

[6] P. Homayounfar and M. L. Owoc, “Data mining research trends in computerized patient records”, Computer Science and Information Systems (FedCSIS), Federated Conference, IEEE, **(2011)**.

[7] N. I. I. Binti and N. H. A. Binti Abdullah, “Developing electronic medical records (EMR) framework for Malaysia's public hospitals”, Humanities, Science and Engineering (CHUSER), IEEE Colloquium, IEEE, **(2011)**.

[8] S. Tsumoto, S. Hirano and Y. Tsumoto, “Information reuse in hospital information systems: A data mining approach”, Information Reuse and Integration (IRI), IEEE International Conference, IEEE, **(2011)**.

[9] J. S. Li, H. Y. Yu and X. G. Zhang, “Data Mining in Hospital Information System”, INTECH Open Access Publisher, **(2011)**.

[10] D. Lo and S. C. Khoo, “Software specification discovery: A new data mining approach”, NSF NGDM, **(2007)**.

[11] S. Tsumoto, “Problems with mining medical data”, Computer Software and Applications Conference, COMPSAC, The 24th Annual International, IEEE, **(2000)**.

[12] M. Arif, “A survey on data warehouse Construction, Processes and Architecture”, International Journal of u- and e- Service, Science and Technology, vol. 8, no. 4, **(2015)**, pp. 9-16.

[13] M. Arif and F. Zaffar, “Challenges in efficient Data warehousing”, International Journal of Grid and Distributed Computing, vol. 8, no. 2, **(2015)**.

[14] M. Arif, K. A. Alam and M. hussain, “Application of Data Mining Using Artificial Neural Network: Survey”, International Journal of Database Theory and Application, vol. 8, no. 1, **(2015)**, pp. 245-270.

## Authors

**Muhammad Arif**, he is a PhD student at Faculty of CS and IT, University of Malaya. Currently he is working on Medical image Processing. His research interests include image processing, E learning, Artificial intelligence and data mining. He joined UM as a Bright Spark Scholar in September 2013 for the period of 3 years. Before this he completed masters and bachelor degrees in Pakistan. He received his BS degree in Computer Science from University of Sargodha, Pakistan in 2011. He obtained his MS degree in Computer Science from COMSATS Islamabad 2013 Pakistan.

**Mehdi Hussain**, he is a Ph.D. candidate at Faculty of Computer Science and Information Technology, University of Malaya, under Faculty development program of National University of Science and Technology (NUST) Islamabad, Pakistan. He received Master and Bachelor of Computer Science degree from SZABIST (2011) and IUB (2005) Pakistan respectively. He served as senior software engineer in Streaming Networks Pvt. (Software House 2006-2014). Research interests are multimedia security, steganography, and data mining. He can be reached at *mehdi141@hotmail.com*.