Clinical Decision Support Using OLAP With Data Mining

Sellappan Palaniappan

Department of Information Technology,
Malaysia University of Science and Technology,
47301 Petaling Jaya, Selangor, Malaysia

Chua Sook Ling

Faculty of Information Technology
Multimedia University,
63100 Cyberjaya, Selangor, Malaysia

Abstract
The healthcare industry collects huge amounts of data which, unfortunately, are not turned into useful information for effective decision making. Decision support systems (DSS) can now use advanced technologies such as On-Line Analytical Processing (OLAP) and data mining to deliver advanced capabilities. This paper presents a prototype clinical decision support system which combines the strengths of both OLAP and data mining. It provides a rich knowledge environment which is not achievable by using OLAP or data mining alone.

Key words:
Healthcare, Clinical decision support system, OLAP, Data mining, hybrid approach

1. Introduction

The healthcare industry is under pressure to lower cost and improve service quality. Oftentimes, information produced is excessive, disjointed, incomplete, inaccurate, in the wrong place, or difficult to make sense [15]. A critical problem facing the industry is the lack of relevant and timely information [10]. As information costs money, it must adopt innovative approaches to attain operational efficiently [2].

Decision Support Systems (DSS) have been developed to overcome these limitations. However, they still do not provide advanced features to help doctors to perform complex queries [2, 17].

Advanced technologies can now generate a rich knowledge environment for effective clinical decision making. This paper presents a prototype clinical decision support system based on OLAP and data mining.

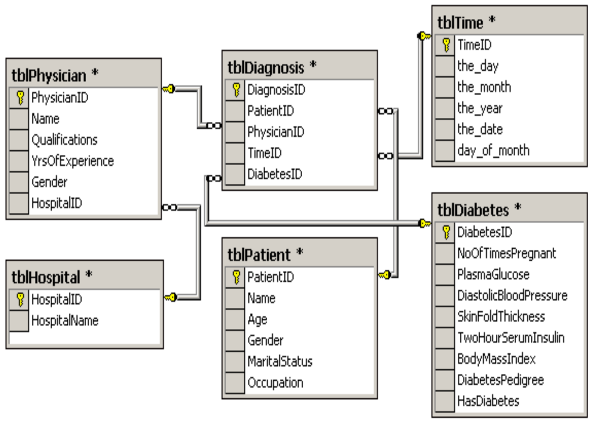
2. Problem Statement

On-Line Transaction Processing (OLTP) systems based on relational databases are suitable for recording business transactions. They record information in two dimensions and automate repetitive tasks. Structured Query Language (SQL) is typically used to access information and results are presented in the form of reports which doctors use to make clinical decisions.

Fig. 1(a) shows a simple OLTP Entity-Relationship Diagram (ERD) consisting six tables and Fig. 1(b) shows a simple SQL query to analyze relationship between hospitals and patients.

OLTP has a major drawback. Large amounts of data in normalized form require many joins even to answer simple queries. For example, to analyze relationships between hospital and patients (Fig. 1), the query would require several table scans and multi-way table joins which can degrade performance significantly [3]. It requires at least four inner joins across five tables (Fig. 2). A real-life database will have many tables and the time taken to process the joins will be unacceptable.

(a)




```
SELECT p.Name, h.HospitalName
FROM tblPatient p, tblHospital h,
     tblTime t, tblDiagnosis d,
     tblPhysician y
WHERE p.PatientID = d.PatientID AND
      d.TimeID = t.TimeID AND
      d.PhysicianID = y.PhysicianID AND
      y.HospitalID = h.HospitalID AND
      t.day_of_month="JAN" AND
      t.day_of_month="JUNE" AND
      h.HospitalName="SJMC"
GROUP BY h.HospitalName
```

Fig.1: (a) ER diagram and (b) SQL command

Manuscript received September 5, 2008.
Manuscript revised September 20, 2008.

(b) **Question:** List the total number of diabetics admitted to "SJMC" in the months of "Jan" and "June"

On-Line Analytical Processing (OLAP) was introduced to overcome this problem. Whereas OLTP uses two-dimensional tables, OLAP uses multidimensional tables called data cubes. OLAP can analyze data in several dimensions. It also incorporates query optimization [3-5]. Users can filter, slice-and-dice, drill-down and roll-up data to search for relevant information efficiently.

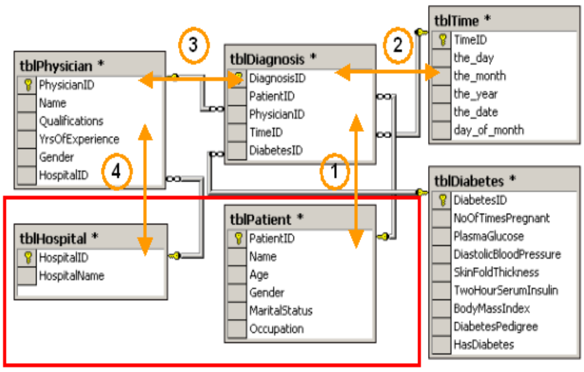


Fig.2: Total number of inner joins involved in analyzing a hospital-patient relationship

However, although OLAP provides an efficient means of analyzing data, it does not identify the dimensions which may be needed to answer a specific decision problem. Also, OLAP cannot predict the future state based on current characteristics.

Fig. 3 shows a simple OLAP cube for six patients under the care of Dr. 20. This information can be used to answer an endocrinologist's question: which of the six patients are likely to be diagnosed with diabetes? OLAP alone cannot answer this question as it does not learn from the data.

Diabetes Diagnosis Details			
HospitalName			
Ampang Puteri Specialist Hospital			
		Year	
		⊞ 1998	Grand Total
Name	Patient Name	TotalDiabetes	TotalDiabetes
⊟ Dr. 20	Bockenkamp Gabriel	1	1
	Brunner June	1	1
	McCoy James	1	1
	Miller Virginia	1	1
	Monitor Alan	1	1
	Sipsy Barbara	1	1
	Total	6	6

Fig. 3: Table representation of data in an OLAP cube

Data mining (defined as a process of nontrivial extraction of implicit, previously unknown, and potentially useful information from the data stored in a database) has now become an important research tool [5-7]. It can discover hidden knowledge in the data and identify problems along several dimensions [7, 23].

However, data mining alone is not very effective. For example, the information given in Fig. 4 does not allow doctors to analyze patients' demographics under category: plasma glucose more than 99.75 and less than or equal to 127.25.
This paper presents a clinical decision support system using OLAP and data mining that can answer complex questions which is not possible by using OLAP or data mining alone.

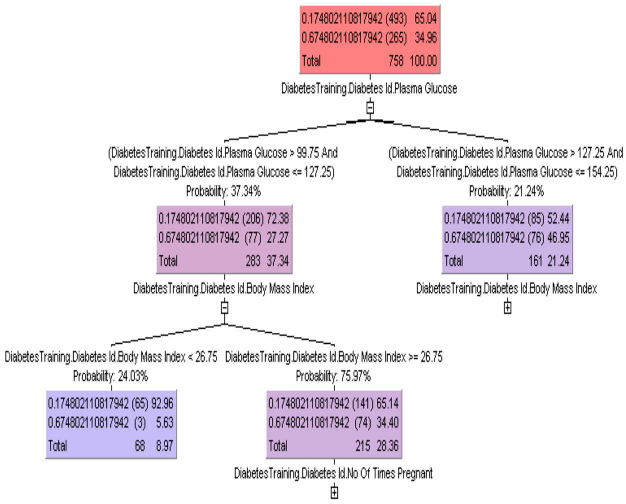


Fig.4: Decision tree mining model

3. Literature Review

Our review covered three areas: On-Line Analytical Processing (OLAP), Application of data mining in healthcare and, Data Mining and Knowledge Discovery in Databases (KDD). It is revealed that organizations generally use OLAP rather than OLTP to build DSS [1, 3-5, 13, 26]. The trend now is to employ combined approaches.

Jonathan et al. [18] used data mining to identify factors contributing to prenatal outcomes as well as quality and cost of prenatal care. Bansal et al. [19] used recurrent neural networks to predict sales forecast for a medical company. Margaret et al. [20] explored the use of artificial neural networks to predict length of hospital stay and its effectiveness in resource allocation. There was very little discussion on the use of data mining in decision support. Hedger [21] used an advanced data mining system to evaluate healthcare utilization costs for GTE's employees and dependents. He also explored the use of IDIS (Intelligence Ware) to discover factors leading to success or failure in back surgeries.

Literature on data warehouse [2, 13, 17], OLAP [1, 14-16], data mining [7, 8, 11] and data warehouse and OLAP integration [3-5] abounds. Parsaye [12] examined the relationship between OLAP and data mining and proposed an architecture integrating OLAP and data mining and discussed the need for different levels of aggregation for data mining. Han [8, 25] explained the need for different levels of aggregation for pattern analysis and focused his work on OLAP mining. His study focused mainly on data mining algorithms, not integration of OLAP and data mining.

4. The Model

As OLAP uses several preprocessing operations such as data cleaning, data transformation, data integration, its output can serve as valuable data for data mining [3, 11].

OLAP operations (e.g., drilling, dicing, slicing, pivoting, filtering) enable users to navigate data flexibly, define relevant data sets, analyze data at different granularities and visualize results in different structures [12, 8, 25]. Applying these operations can make data mining more exploratory.

The motivation for an integrated model, OLAP with data mining, is the concept hierarchy. Data in OLAP and decision tree are organized into multiple dimensions where each dimension contains multiple levels of abstraction defined by the concept hierarchy [8, 29]. The concept hierarchy is illustrated in Fig. 5, where each member has one root and all members between roots have parents and every branch ends with a leaf member.

OLAP data cubes which store concept hierarchies can be used to induce decision trees at different levels of abstraction [29, 1]. Once the decision tree mining model is built, the concept hierarchies can be used to generalize individual nodes in the tree, which can then be accessed by OLAP operations and viewed at different levels of abstraction.

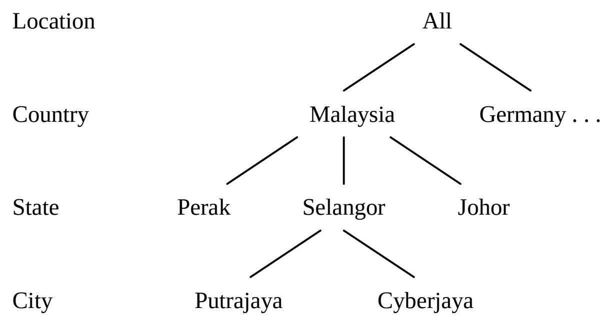


Fig. 5: A concept hierarchy for the dimension location

5. Research Questions

This research demonstrates how the integrated approach, OLAP with data mining, provides advanced decision support compared to using OLAP or data mining alone. The research questions listed in Table 1 are used for this purpose. They cannot be answered by using OLAP or data mining alone, but can be answered by using the integrated model.

Fig. 6 shows the architecture of the integrated model (OLAP with data mining) comprising of several components. The system is divided into two parts: *Server-side* – for building the integrated model, and *Client-side* – for accessing queries and presenting results (Fig. 6). It uses OLAP operations (slice, dice, roll-up, drill-down, pivot) and the decision tree mining algorithm C4.5. The test data validates the effectiveness of the model.

Table 1: Research Questions

R1.	How does OLAP with data mining *enhance real time indicators* like bottlenecks?
R2.	How does OLAP with data mining provide *improved visualization* to uncover patterns/trends that are likely to be missed?
R3.	How does OLAP with data mining *uncover more subtle patterns* in data over capabilities provided by OLAP or data mining alone?

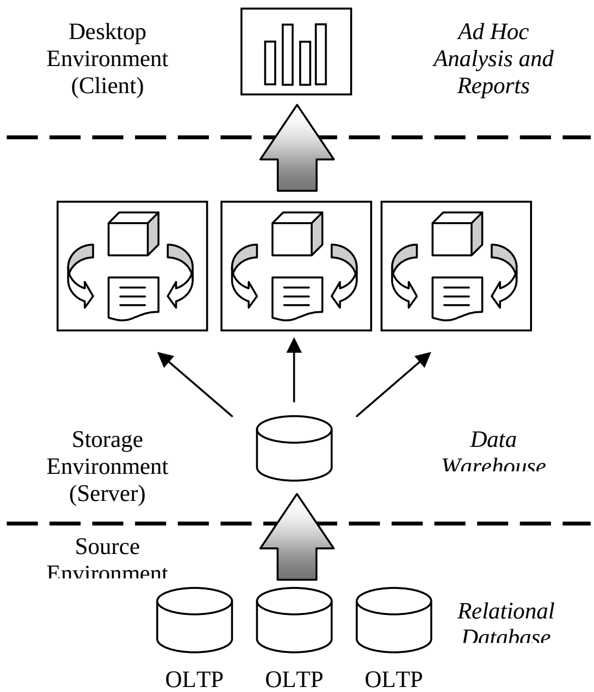


Fig.6: Integration of OLAP with data mining architecture

6. System Implementation

A data cube is first created then the data mining process is started. The cube preserves the information and allows browsing at different conceptual levels. It serves as the data source for the data mining task. Data mining can be performed on any level or dimension of the cube. After the model is built it is stored in the OLAP cube. Each dimension represents the rule corresponding to a node in the decision tree mining model (Fig.7). OLAP operations explain the different states of the system.

The data for this study is taken from UCI Repository of Machine Learning Databases [9]. The data comprise of *Pima Indian Diabetes* database, *Post Operative Recovery* database and *Liver Disorder* database. As the data is declassified, we have added several dummy attributes such as patient, doctor and hospital information.

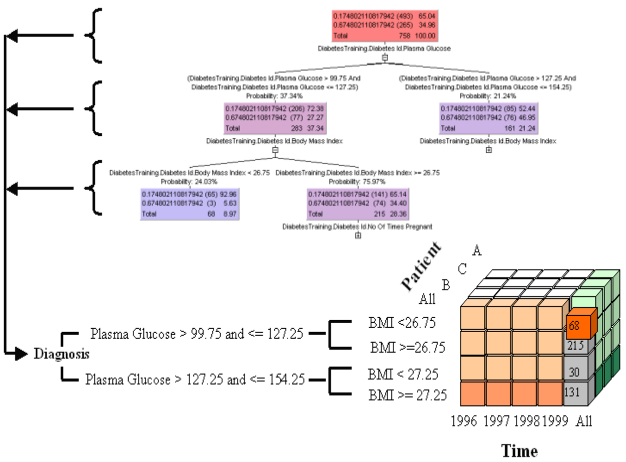


Fig.7: A logical view of representation of decision tree mining model in an OLAP cube

A two-year sample dataset (1997-1998) is created to mine for knowledge discovery. Information on entities and their attributes and relationships are fed into the data warehouse. Its design is based on the star schema (Fig. 8).

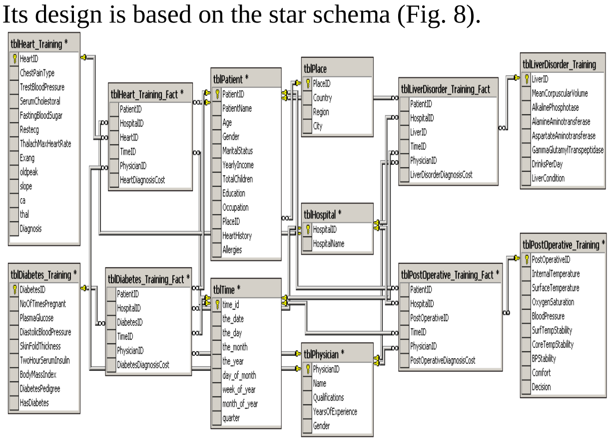


Fig.8: Data warehouse star schema design

7. System Implementation

The system can predict the future state and generate useful information for effective decision-making. It can answer all the research questions listed in Table 1.

R1. The integrated model enhances real time indicators by using information on hospital room utilization for post-operative recovery patients. It allows hospital administrators to discover any bottlenecks that might exist. It allows them to solve problems related to hospital room utilization. The results in Fig. 9 show that a total of 6 patients are likely to be warded in Hospital A. The administrator can use this information to allocate rooms based on their characteristics. For patients over 60 years, a decision may be made to ward them in senior citizen's ward or transfer them to another hospital. This indicator is useful for performing "what-if" analysis on hospital room availability.

R2. The integrated model improves information visualization. It discovers overall trends that are likely to be missed by using OLAP or data mining alone. Fig. 10(a) shows that of all the states in Malaysia, Perak has the most number of patients diagnosed with liver disorders. Barring this, the best liver disorder specialists are from Kuala Lumpur (Fig. 10(b)). It provides a more comprehensive analysis and facilitates decision-making by allocating physicians to under-represented geographical areas. It allows the quality of physicians in under-represented areas to be improved.

R3. With data mining, doctors can predict patients who might be diagnosed with diabetes. OLAP provides a focused answer using historical data. However, by combining, we can optimize existing processes and uncover more subtle patterns, for example, by analyzing patients' demographics.

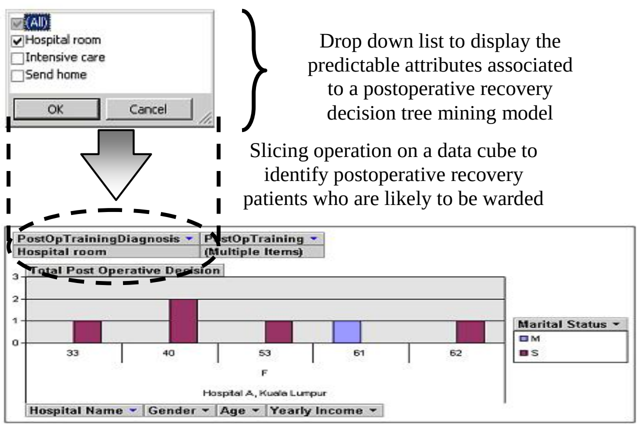
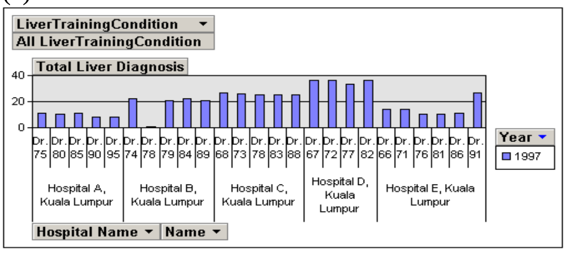


Fig.9: The results of prediction to identify postoperative recovery patients that are likely to be warded at Hospital A

Fig. 11 shows that patients whose plasma glucose is more than 154.25 have a higher probability of being diagnosed with diabetes compared to those patients whose plasma glucose is between 127.25 and 154.25. It also shows that female patients who are single are more likely to be diagnosed with diabetes compared to male patients.

Interestingly, for patients who have reached a certain stage in the process of care, doctors (endocrinologists) can recommend that they undergo retinal examination (Fig. 12). This can help to reduce or prevent blindness before reaching a critical stage.

(a)



(b)

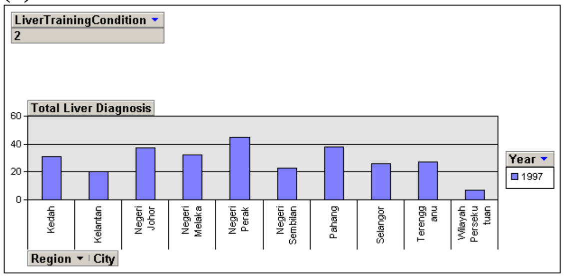
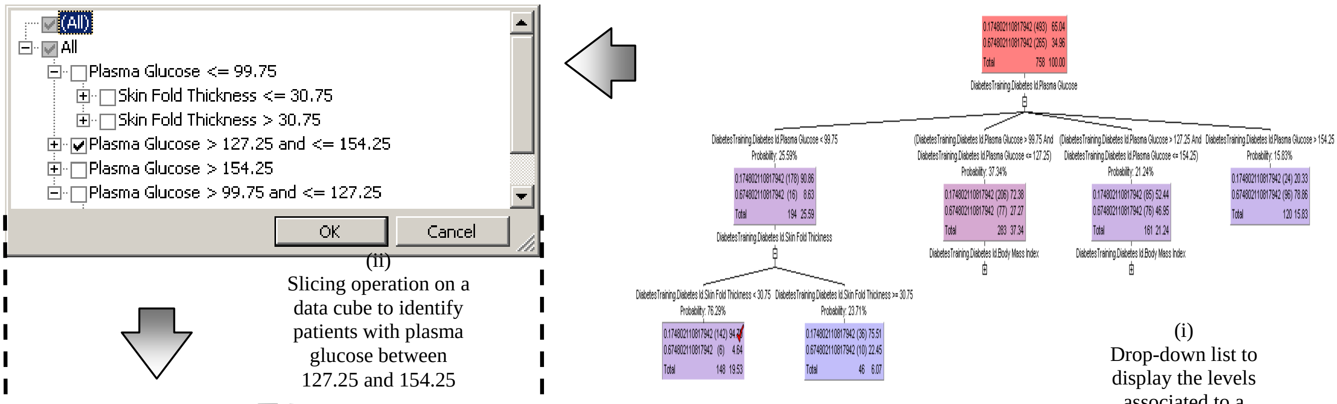


Fig.10: (a) The results of prediction to identify patients who may be diagnosed with liver disorders broken down by regions
(b) OLAP analysis to identify physicians specializing in liver disorders

The system allows users to perform advanced data analyses and ad hoc queries. It also generates reports in multiple formats.



Diabetes Diagnosis Details

DiabetesTrainingDimension	DiabetesTrainCondition	
Plasma Glucose > 127.25 and <= 154.25	1	
	Gender	
	F	Grand Total
Marital Status	Total Diabetes Diagnosis	Total Diabetes Diagnosis
M	32	32
S	44	44
Grand Total	76	76

(a)

Diabetes Diagnosis Details

DiabetesTrainingDimension	DiabetesTrainCondition	
Plasma Glucose > 154.25	1	
	Gender	
	F	Grand Total
Marital Status	Total Diabetes Diagnosis	Total Diabetes Diagnosis
M	46	46
S	50	50
Grand Total	96	96

(b)

Fig.11: Comparison results between patients whose plasma glucose (a) between 127.25 and 154.25 and (b) plasma glucose more than 154.25

8. Conclusion and Future Work

This paper has presented a DSS based on OLAP with data mining. The system is powerful because (1) it discovers hidden patterns in the data, (2) it enhances real-time indicators and discovers bottlenecks and (3) it improves information visualization.

Further work can be done to enhance the system. For example, features can be added to allow doctors to query data cubes on business questions and automatically translate these questions to Multi Dimensional eXpression (MDX) queries. The model can also include complex data objects, spatial data and multimedia data. Besides decision tree, the use of other data mining techniques can also be explored.

Acknowledgements

This paper was supported by E-Science grant (No: 01-02-05-SF0017), Ministry of Science and Technology and Innovation (MOSTI), Government of Malaysia.

References

[1] Helen, H. and Peter, H., Using OLAP and Multidimensional Data for Decision Making, IEEE IT Professional, 44-50, 2001, October.

[2] Donald, J.B., John, W.F., Alan, R.H., James, S., Healthcare Data Warehousing and Quality Assurance, IEEE Computer, 56-65, 2001, December.

[3] Surajit, C. and Umeshwar, D., An Overview of Data Warehousing and OLAP Technology, ACM Sigmod Record, *26*(1), 65-74, 1997.

[4] Ralph, K. and Margy, R., The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling (2nd ed.), Canada: John Wiley & Sons, Inc, 2002.

[5] Torben, B.P. and Christian, S.J., Multidimensional Database Technology, IEEE Computer, *34*(12), 40-46, 2001, December.

[6] Usama, M. F., Data Mining and Knowledge Discovery: Making Sense Out of Data, IEEE Expert, 20-25, 1996, October.

[7] Ming-Syan, C., Jiawei, H. and Philip, S.Y., Data Mining: An Overview From a Database Perspective, IEEE Transactions on Knowledge and Data Engineering, *8*(6), 866-883, 1996, December.

[8] Han, J., OLAP Mining: An Integration of OLAP with Data Mining, Proceedings of 1997 IFIP Conference on Data Semantics (DS-7), Leysin, Switzerland, 1-11, 1997, October.

[9] Blake, C.L. & Merz, C.J., UCI Repository of Machine Learning Databases, University of California, Department of Information and Computer Science, 1998. Available: http://www.ics.uci.edu/~mlearn/ MLRepository.html

[10] Sellappan, P., Chua, S.L., Ng, Y.H., Ng, Y.M., Healthcare Information Services - The Application Service Provider (ASP) Model, Proceedings of SEARCC Conference 2004, Kuala Lumpur, October, 2004.

[11] Fayyad, U., Gregory, P.-S. and Smyth, P., From Data Mining to Knowledge Discovery in Databases, AI Magazine, *37*(3), 37-54, 1996.

[12] Parseye, K., OLAP and Data Mining: Bridging the Gap. Database Programming and Design, *10*, 30-37, 1998.

[13] Surajit, C., Umeshwar, D., and Ganti, V., Database Technology for Decision Support Systems, IEEE Computer, *34*(12), 48-55, Dec. 2001.

[14] Panos, V., and Timos, S., A Survey on Logical Models for OLAP Databases. ACM Sigmod Record, *28*(4), 64-69, Dec. 1999.

[15] Robert, S.C., Joseph, A.V. and David, B., Microsoft Data Warehousing: Building Distributed Decision Support Systems, London: Idea Group Publishing, 1999.

[16] George, C., OLAP, Relational and Multidimensional Database Systems, Acm Sigmod Record, *25*(30), 64-69, Sept. 1996.

[17] Shim, J.P., Warkentin, M., Courtney, J.F., Power, D.J., Ramesh, S., and Christer, C., Past, Present and Future of Decision Support Technology, Elsevier Science B. V., *33*, 111-126, 2002.

[18] Jonathan, C.P., Lobach, D.F., Goodwin, L.K., Hales, J.W., Hage, M.L. and Hammond, W.E., Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse, Proceedings of the American Medical Informatics Association Symposium, Philadelphia, United States of America, 101-105, 1997.

[19] Bansal, K., Vadhavkar, S., and Gupta, A., Neural Networks Based Data Mining Applications for Medical Inventory Problems, International Journal of Agile Manufacturing, *2*(1), 187-200, 1998.

[20] Margaret, R.K., Kevin, C.D., and Ida, A., Data Mining in Healthcare Information Systems: Case Study of a Veterans' Administration Spinal Cord Injury Population, Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03), Hawaii, United States of America, IEEE Computer, 159-167, 2002.

[21] Hedger, S.R., The Data Gold Rush, Byte, *20*(10), 83-88, 1995.

[22] Bill, G. F., Huigang, L. and Kem, P. K., Data Mining for the Health System Pharmacist. Hospital Pharmacy, *38*(9), 845-850, 2003.

[23] Usama F., Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases. Proceedings of the 9th International Conference on Scientific and Statistical Database Management (SSDBM ’97), Olympia, WA., 2-11, 1997.

[24] Raymond P.D., Knowledge Management as a Precursor Achieving Successful Information Systems in Complex Environments. Proceedings of SEARCC Conference 2004, 127-134, Kuala Lumpur, Malaysia.

[25] Han, J., Chiang, J.Y., Chee, S., Chen, J., Chen, Q., Cheng, S. & et al., DBMiner: A System for Data Mining in Relational Databases and Data Warehouses, Proceedings of the 1997 Conference of the Centre for Advanced Studies on Collaborative research, Ontario, Canada, 1-12, November, 1997.

[26] Sarwagi, S., Explaining Differences in Multidimensional Aggregate, Proceedings of the 25th International Conference on Very Large Data Bases, Scotland, United Kingdom, 42-53, September, 1999.

[27] Fong, A.C.M, Hui, S.C., and Jha, G., Data Mining for Decision Support, IEEE IT Professional, *4*(2), 9-17, March/April, 2002.
[28] David K. and Daniel O'Leary, Intelligent Executive Information Systems. IEEE Expert, *11*(6), 30-35, Dec. 1996.
[29] Han, J., Kamber, M., Data Mining Concepts and Techniques, San Diego, USA: Morgan Kaufmann Publishers, pp. 294-296.

Sellappan Palaniappan obtained his PhD in Interdisciplinary Information Science from University of Pittsburgh and a MSc in Computer Science from University of London. He is an Associate Professor at the Department of Information Technology, Malaysia University of Science and Technology. His research interests include information integration, clinical decision support systems, OLAP and data mining, web services and collaborative CASE tools.

Chua Sook Ling obtained her MSc in Information Technology from Malaysia University of Science and Technology (MUST). She is a Lecturer in the Faculty of Information Technology, Multimedia University (MMU). Her research interests include data mining, OLAP, business intelligence, decision support systems and knowledge management.