

A Proposed DDS Enabled Model for Data Warehouses with Real Time Updates

Munesh Chandra Trivedi¹, Virendra Kumar Yadav^{*2}, Avadhesh Kumar Gupta³

^{1,2}Department of Computer Science and Engineering, ABES Engineering College, Ghaziabad, Uttar Pradesh, India

³Departement of Information Technology, IMS Ghaziabad, Uttar Pradesh, India

Article Info

Article history:

Received Jan 18, 2018

Revised Mar 6, 2018

Accepted Mar 13, 2018

Keyword:

Deviation detection system
(DDS)

Etl

Mapping engine

Triggered update

ABSTRACT

Data warehouse generally contains both types of data i.e. historical & current data from various data sources. Data warehouse in world of computing can be defined as system created for analysis and reporting of these both types of data. These analysis report is then used by an organization to make decisions which helps them in their growth. Construction of data warehouse appears to be simple, collection of data from data sources into one place (after extraction, transform and loading). But construction involves several issues such as inconsistent data, logic conflicts, user acceptance, cost, quality, security, stake holder's contradictions, REST alignment etc. These issues need to be overcome otherwise will lead to unfortunate consequences affecting the organization growth. Proposed model tries to solve these issues such as REST alignment, stake holder's contradiction etc. by involving experts of various domains such as technical, analytical, decision makers, management representatives etc. during initialization phase to better understand the requirements and mapping these requirements to data sources during design phase of data warehouse.

Copyright © 2018 Institute of Advanced Engineering and Science.

All rights reserved.

Corresponding Author:

Virendra Kumar Yadav,
Department of Computer Science and Engineering,
ABES Engineering College,
Ghaziabad, Uttar Pradesh, India 201009, India.
Email: virendrashines@gmail.com

1. INTRODUCTION

In the today's world of internet era, role of electronic information cannot be ignored. Effective processing of these electronic information helps manager in daily activities during decision. According to Inmon, data warehouse can be defined as subject oriented, nonvolatile, time-varying and integrated collection of data and these attributes makes its different from operational databases [1]. Data warehouse enables its users generally decision makers such as managers to make better and faster decisions during strategic planning. If data provided by data warehouse is efficient and informative then it helps managers to take better decision during the strategic planning of an enterprise. Need of automated warehouse is also increasing now a days. If automation of warehouses is done correctly, it helps in reducing costs, efforts and the most important to reduce the human errors which may lead to incorrect decisions corresponding leads to inefficient utilization of data warehouse. Several researchers concluded ERP as solution. ERP stands for Enterprise Resource Planning. With the help of ERP, an enterprise can perform certain operation of enterprise such as collect, store, interpret and manages data from many processes. ERP if successfully implemented will results in efficient utilization of resource and efficient management which leads to better decisions. Implementing ERP requires significant investment. If it fails it may lead to significant financial losses. Challenges that were identified in literature are business process complexities, proper understanding the organization needs, skilled staff etc. Around 66 to 70

organizations which implemented ERP were failed to reap the benefit the ERP [2]. Data warehouse can be think as one of the important components without which existence of decision support system is very hard to realize. Now a day's policy makes or decision makers of an enterprise started deploying and focusing on maintaining Data warehouse. Any industry for example let's say retail, telecommunication, healthcare manufacturing etc. are maintaining Data ware houses to support their decisions or simple words to improve their decision making capabilities [1][5]. This improved decision making helps in accuracy in future forecast or to improve their earnings in business.

Preparation of Data warehouse passes through three major operations i.e. Extract, Transform and Loading. In short these processes are combining known as ETL. Collection of data from different data bases may contains errors and anomalies. If such data is directly put into data warehouse and decision is taken on this prepared data ware house, then? Definitely it will results in wrong reflection in output or organization performance. So all these three processes i.e. ETL needs to perform to clean the data and then loading these data to data warehouse is performed.

Rest of the paper is organized into five sections. Section II contains the literature survey. Section III contains the proposed method. Section IV contains the proposed algorithm and expected outcomes. Section V contains the conclusion and future research directions.

2. SYSTEMATIC LITERATURE REVIEW PROCESS

Authors in their paper conducted a survey which was based on Telecommunication Company. Telecommunication Company has small warehouse consisting of scratch cards and simcards. The whole process is carried using manual entry excel sheets. The aim of this survey is to find out the processes or procedures which can be automated. When this step is completed successfully, another step is to choose software program. Software program is chosen according to need of an enterprise and can withstand with the large amount of data. Automation of warehouse helps in controlling, movement and storage of products along with enhanced security. Author in their automation applied the FIFO concept.

Authors Nur Hani et al. in their paper titled "User Requirement Analysis in Data Warehouse Design: A Review" discussed about the various analysis approaches that focusses on the role of user requirement in data warehouse design. Four broad categories in which user requirements approaches can be classified are: Goal driven, data driven, mixed driven and mixed approaches. These classification was performed by the researchers in order to identify the role of user requirements but it is very difficult for data warehouse designer to find out the suitable technique which they should select in designing of data warehouse [4]. Author in this paper also discuss about the strength and weakness in these four categories. According to the author the most critical phase in data warehouse development is requirement analysis. In the papers [6] [7] author also shown that 80% DW project fail to fulfil business objectives. Because of variation in end user. Some of the researcher also mentioned in their papers about ignorant behavior of decision makers towards this phase [9] i.e. Requirement analysis phase. They were more concerned about technical aspects rather than requirement analysis phase [8]. Concept is more clear if any IT people will unable to understand or there is miscommunication between IT and policy makers or decision makes, will lead to poor data warehouse design which ultimately results in failure of data warehouse objectives [10].

The first approach is Data-driven approach. Some research papers refers this Data-driven approach by other name known as supply-driven approach [1] [7]. In this kind of approach Database administrator plays a very important role. Transactional data is analyzed and logical schema is build. Generally this kind of approach eliminates the need of user involvement.

Second approach is user driven approach [11]. This approach uses the concept of bottom up. Project manager plays a key role. Project manager has the responsibility to document all the requirements of different business user. This documented information is integrated with data warehouse.

Goal-Driven is third approach. In this approach top level management plays an important role. The management person or policy makers decides the goal priorities. Based on these goals, data warehouse is expected to give the answers i.e. how much these goals have been achieved [12].

Fourth approach is Mixed-Driven approach. This kind of approach have been develop to strengthen the requirement analysis.

Winter & Strauch [13] in their research paper proposed an approach that requires two things, identify the end users that plays lead role in decision making in an organization and an application that can connect data warehouse to information. End user will decide their organizational requirements but in priority wise. Before these requirements is finally converted into information and finally mapped with data warehouse, this end user requirement process is iterated till end user satisfies with its outcome.

Data driven approach has several strengths such as data availability decides design of data warehouse. The schema generated with this approach is known for their stability [4]. But this approach almost ignore the

involvement of end user. Also some of the researcher agrees on this point that it is very difficult to perform the ETL process on large data sources in order to generate relevant information. In user driven approach, end user gets priority. This kind of approach is highly appreciated by the end user. But this kind of approach has certain limitations. Such as it is very difficult to satisfy all the requirements of end user by mapping it with warehouse.

Authors agrees on this point that requirement engineering must be performed to ensure the smooth process. Shao et al [14] in their research talked about the Real-time data warehouse. In their research they researched about the structure of real time data warehouse. They structured the data warehouse which is based on double mirror replication mechanism and multi-level caches.

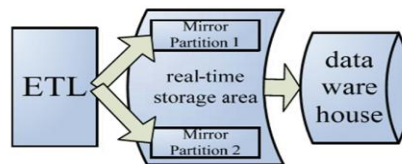


Figure 1. Real-time Storage Area Structure [14]

Now a days the data ware house cannot be considered in isolation for decision making. Competitive world of today's is demanding. To make policies/decisions, both data i.e. organization data & data from outside worlds (competitors) are required. Security is also measure concerned now a day's. Use of good encryption technique/algorithms can be a solution (old solution). Authors in their paper compared the various techniques proposed in various articles on the basis of securities parameters such as: Encrypted data, Audit control, extendibility, platform independence model security, transformation, creation of PSM, QVT support, integration of multiplatform data.

Table 1. Mixed Data Driven Approaches [4]

Articles	E D	A C	E x	C I	P I	T r	P S	QV T	I	I D	P	Cit	IF	SC
[Caserta & Kimball, 1998]	Y	N	N	N	N	N	N	Y	Y	N	Wiley	959	NA	In-stream Encryption
[Kirkgoze et al., 1997]	N	N	N	N	N	N	N	Y	N	N	IEEE Conf.	60	NA	Role Based Access Profile Meta Data
[Katie et al., 1998]	N	N	Y	N	N	N	N	Y	N	N	IEEE Conf.	86	NA	UML Extension
[Eduardo et al., 2004]	N	N	Y	N	Y	N	N	N	N	N	LNCS Springer	26	NA	UML2.0 + CWM
[Emilio et al., 2006]	N	Y	Y	N	Y	N	N	N	N	N	LNCS Springer	4	NA	UML2.0/ OCL Extension UML Ext.
[Eduardo et al., 2006]	N	Y	Y	N	Y	N	N	N	N	N	JRPIT	30	0.22	MDA + QVT + CWM
[Eduardo et al., 2007]	N	N	Y	N	Y	N	Y	N	Y	N	Elsevier	66	1.235	MDA + QVT + SSAS
[Jose & Juan, 2008]	N	N	Y	Y	Y	S	Y	Y	N	N	Elsevier	159	2.036	UML 2.0 + MDA + ADM
[Emilio et al., 2008]	N	N	Y	N	Y	S	Y	Y	Y	N	Elsevier	20	1.177	MDA + SPEM
[Carlos et al., 2008]	N	N	Y	Y	Y	F	Y	Y	Y	N	LNCS Springer	7	NA	
[Emilio et al., 2009]	N	Y	Y	Y	Y	N	Y	N	Y	N	Elsevier	14	1.177	
[Amulfo et al., 2009]	N	Y	Y	Y	Y	F	Y	Y	N	N	LNCS Springer	6	NA	
[Emilio et al., 2009]	N	Y	Y	Y	Y	S	N	Y	N	N	Elsevier	23	1.328	

There is certain issues which must be taken care while designing of data warehouse. Data warehouses are decisional information artifacts that are embedded in the organizations that create/maintain them. Therefore, their contents must be highly supportive of the decision-making activity of organizations. The decision-making activity, in turn, is tightly coupled to the goals that an organization sets for itself. But the approaches discussed above do not take into account the larger organizational context in which the DW is to function.

- a. How can we ensure correct requirements? Correct query set that data warehouse is supposed to answer.
- b. Re-examine the notions of goals and scenarios for data-oriented systems.
- c. It can be seen that the requirements engineering problem for data warehouse systems is the inverse of that for functional systems, the former is aimed at the discovery of data and de-emphasizes functionality whereas the latter aims to discover the functionality of systems and de-emphasizes data discovery. This shift in emphasis demands for re-examination of the notions of goals and scenarios for data warehouse systems.
- d. How can an actor (stake holders) ensure about facts that are provided by data warehouse are meeting with the expectation in decision making process and in their success?
- e. Looking at software engineering and information system view so requirements engineering in context of Data Warehouse. It is well known that a data warehouse can be looked upon from the organizational and from the technical perspectives. The former looks upon the warehouse as embedded in an organization and considers the manner in which it supports organizational tasks. The latter deals with issues of data warehouse contents, their structure etc. The organizational view of data warehouse corresponds to the Information Systems perspective of Requirements Engineering whereas the technical view corresponds to the Software Engineering view. None of the approach for data warehouse development discusses the development of data warehouse from both points of view.
- f. Who should involve in requirement identification phase?
- g. How to avoid contradiction between expectations of various stakeholders and designed data warehouse?
- h. General lack of specific guidance for the requirement elicitation process for Identification of data warehouse contents. Number of authors has proposed to adapt traditional requirements engineering approach in specific context of development of data warehouse. But these approaches lack in specific guidance for requirements elicitation [15], [16], [17]. For example, the proposal of [Fab03] to build a framework for DW requirements engineering provides pointers to RE approaches that may be applicable, but does not establish their feasibility and also does not consider any detailed technical solutions.
- i. Lack of Automation of the Requirements Elicitation Process. None of the approach provides automation of the application of the requirements elicitation process. Few CASE tools for DW conceptual design have been implemented. In ADAPT and in GOLD, conceptual schema is directly drawn by the designer but no active support for requirements elicitation is provided.

3. PROPOSED WORK

This section discusses about the proposed work.

3.1. Initialization Phase

The first step is to identify the correct expectations from data warehouse of an actor (actor can be business experts, analyst experts, stake holders, project managers etc.). To find out the correct expectations to ensure correct decision, concept of a formal discussion (which can take place through online or) is proposed. It is expected that all actors such as business experts, analyst experts, stake holders, project managers etc. should be part of this discussion phase. All the experts (includes business experts, analyst experts, stake holders, project managers and any other important management or decision making person) will put their expectations (in form of draft document). Now this draft document will be verified by the technical experts (software engineer, DBA etc.) to ensure the valid expectations from designed data warehouse prepared from various data sources. If technical experts team find some invalid expectations or say some expectations for which out data sources doesn't contains any supportive facts will considered for elimination otherwise final draft is prepared and send or informed to every experts involved in discussion phase. Final draft is actually query sets which designed data warehouse software is expected to answer.

Mapping engine will contains program which is designed by the software team in order to map the query set requirements to data sources in order to create data warehouse. Mapping engine will also contains an intelligent program DDS which is responsible for triggered update.

3.2. Update Phase

Once when the first phase is completed successfully i.e. requirements or query sets is mapped with data sources and finally data warehouse is built, now it is ready for its users to ask queries and providing them

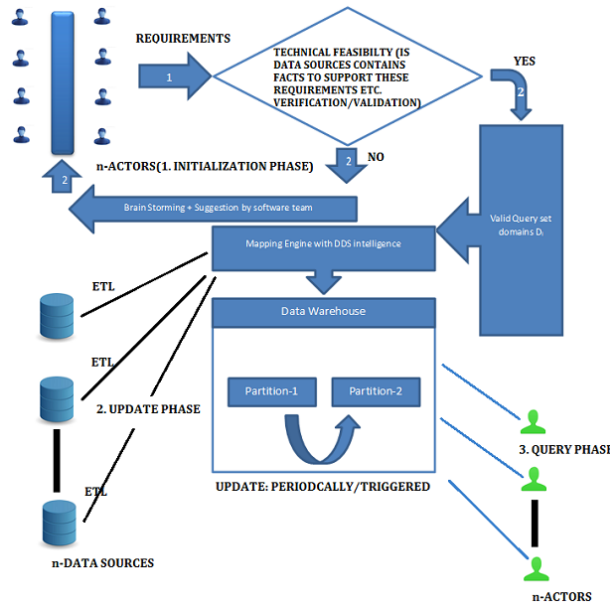


Figure 2. Deviation Detection System (Proposed Approach)

ETL: E-Extraction, T-Transform, L-Loading

Actors: Can be business experts, analysis experts, stakeholders, requirement engineers, project managers etc.

Update: Periodically/Triggered \rightarrow from Partition-1 to Partition-2

Query set domains: $D_i \{D_1(Q^1_1, Q^1_2 \dots Q^n_n), D_2(Q^2_1, Q^2_2 \dots Q^2_n) \dots D_n(Q^n_1, Q^n_2 \dots Q^n_n)\}$ and their respective query set Q^i_j where $i = 1, 2, 3 \dots n$ & $j = 1, 2, 3 \dots n$.

Verify and validate query domain D_i and their query set Q^i_j to filter query domains for their correctness, completeness, consistency and to meet expectation of stakeholders/decision makers.

Accurate answers. But what about update. As every time data sources are receiving records and these records after the ETL process should be loaded to Data warehouses to ensure accuracy in decision making. Proposed model includes two types of updates i.e. periodically and triggered update. Virtually data warehouse is divided into two parts: partition-1 and partition-2. Partition-1 contains the current time records which was uploaded to data warehouse by the mapping engine from data sources after performing ETL process. Partition-2 contains the historical records or records up to a certain period (i.e. information before periodic update can take place).

Mapping engine will update the data warehouse partition-1 after a period of time as defined in mapping engine program. There is also a provision of triggered updates. This kind of update takes place when DDS detects deviation from the expected pattern. DDS which is an intelligent program embedded in mapping engine will continuously monitoring the pattern from data sources. When DDS detects deviation $\geq T_{dev}$, will set the FLAG == Triggered Update which will result in immediate update of partition-2 from partition-1 and simultaneously an alert is generated which is send to its users to catch their attention.

4. PROPOSED WORK

The Proposed algorithm is as follows:

4.1. Initialization Phase

- Define Query set domains $D_i \{D_1(Q^1_1, Q^1_2 \dots Q^n_n), D_2(Q^2_1, Q^2_2 \dots Q^2_n) \dots D_n(Q^n_1, Q^n_2 \dots Q^n_n)\}$ and their respective query set Q^i_j where $i = 1, 2, 3 \dots n$ & $j = 1, 2, 3 \dots n$.
- Verify and validate query domain D_i and their query set Q^i_j to filter query domains for their correctness, completeness, consistency and to meet expectation of stakeholders/decision makers. i.e. if R_i is set of queries that needs to be filtered out from domain D_i then

$$\text{Final_query_domain} = \{D_i(Q^i_j) - R_i(Q^i_j)\}$$
- Perform ETL through Mapping Engine.
- Load Data_(ETL)(from data source) into DW_partition².

4.2. Update Phase (DW_Partition¹ → DW_Partition²)

This phase occur under two circumstances:

A. Periodic Update:

UPDATE

1. Mapping engine -----→ DW_partition¹.
(Only new record)
2. If (T_period ≥ T_{threshold}) // T_period is time period after which DW will update.

UPDATE

Mapping engine -----→ DW_partition²
(Only new record)

3. if (T_period < T_{threshold})

DW_partition¹ -----→ DO NOTHING.

B. Triggered Update:

1. mtr_Mapping_Engine (N_Data Sources, DW_Partition²) // mtr_Mapping_Engine is monitoring agent in mapping engine.

if (dev(obs_N data sources)_pattern – dev(data_DW_Partition² pattern) ≥ T_{dev})

// obs_N data sources: Observed pattern generated through new current records in data sources.

// T_{dev}: Decided by organization major role playing actors.

```
{
  Raise alarm
  SEND ALERT to concerned person
  DW_partition1 -----→ DW_partition2
}
else
DO NOTHING
```

4.3 Query Phase

A. Authentication

login(Actor_X, user_ID, PWD)

// PWD: password

// Actor_X: Concerned person

if (entered_user_ID == Recorded_user_ID && entered_user_PWD == Recorded_PWD)

```
{
  message "authentication successful";
  message "ask queries" from DW_Partition2;
}
else
{
  message "authentication unsuccessful" or "try again"
  go back login_screen;}
```

5. RESULTS AND OBSERVATIONS

Here are some of the queries (issues) which proposed model is able to answer.

Query 1) How to ensure correct requirements to meet expectations of every actor?

Verification and validation procedure in requirements/expectation during initialization phase.

Query 2) How to avoid contradiction between expectations of various stakeholders and designed data warehouse?

By involving database designer and software experts in initialization phase. Involving database designer and software experts in initialization phase will ensure verification of correct requirements i.e. approx. correct mapping b/w their query set into required data warehouse to meet expectation of their user.

Query 3) How it improves REST alignment?

If REST alignment is not done in an efficient manner, it will lead to defective development of data warehouse or simply efforts will be wasted. Misalignment will lead to disappointment as what the experts are expecting

from data warehouse software, is unable to answer or convince or to provide supportive facts through which forecast or decision could be taken. Involving participation of management experts, top officials, analyst experts, advisors or any other experts along with technical experts during the initialization phase will ensure approx. accurate alignments.

Query 4) Does the proposed model include provision for real time data warehouse update?

Yes, proposed model includes two types of updates i.e. periodic update $\geq T_{\text{period}}$ & Triggered Update

When mapping engine module (inbuilt with deviation detection system (DDS)) detect some pattern which is not expected, it will immediately raise an alarm and update flag will be generated and data warehouse is updated immediately.

Table 2. Comparisons of Mixed Data Driven Approaches [4] with Proposed Approach

Mixed Approaches	GRAnD [13]	CADWA [3]	Mazon et al. [10]	Jukic & Nicolas [14]	Winter & Strauch [8]	IPD [4]	Triple Driven Approach [1]	DWARF [12]	Proposed Approach
Top Management Involvement	High	High	High	Low	Low	Moderate	High	High	High
End User Involvement	Low	High	High	High	High	High	High	Moderate	High
Notation / Diagram	Adopt Tropos notation	Not define	Adapt i* technique	Not Define	Not Define	Adapt ERM Model	Not define	Adapt UML	Not define
Time Consume Modeling	Moderate	High	Moderate	High	Moderate	High	Moderate	Moderate	Moderate
Technique Complexity	High	High	Low	Moderate	High	High	Low	Moderate	Low
Fulfill business goal	High	High	Moderate	Low	Low	Moderate	High	High	High
Schema Quality	High	High	High	Moderate	Moderate	Moderate	High	Moderate	High

6. CONCLUSION

Increasing dependency on digital world gives birth to role of electronic information. If processing of electronic information is done effectively, helps in decision making and better future forecast. Data warehouse construction involves certain issues which needs to be resolved or minimized in case if they are difficult to eliminate. Issues may arise due to inconsistency of data, conflicts between logic, cost, user acceptance REST alignment etc. In this paper an approach named Deviation Detection System (DDS) has proposed. DDS approach tries to solve these above mentioned issues upto certain extent (may vary organization to organization needs but improved observations can be made). From table 2. It can be clearly observed that proposed algorithm reflect improved observations.

ACKNOWLEDGEMENTS

I would like to thanks my professors who of course is expert in this area, for their valuable suggestions and discussion hours. I would like to thanks my university management for providing me the resources which may be in the form of journals or internet or financial grants.

REFERENCES

- [1] Inmon WH. "Building the data warehouse (second edition)". John Wiley and Sons; 1996.
- [2] Haddara, M. "ERP Selection: The SMART Way," *Procedia Technol.*, 2014;16: 394–403.
- [3] Anas M. Atieh, Hazem Kaylani, Yousef Al-abdallat, Abeer Qaderi, Luma Ghoul, Lina Jaradat, Iman Hdairis. "Performance improvement of inventory management system processes by an automated warehouse management system". Published in 48th CIRP Conference on Manufacturing Systems-CIRP CMS 2015: 568 – 572.

- [4] Nur Hani ZulkifliAbaia, Jamaiah H. Yahayab, Aziz Deramanc. “*User Requirement Analysis in Data Warehouse Design: A Review*”. Published in proceedings of 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013), 801 – 806.
- [5] W. H. Inmon. Building the Data Warehouse. 5th Edition. John Wiley & Sons; 2005.
- [6] J.-N. Mazon, J. Trujillo, M. Serrano, M. Piattini. Designing Data Warehouses: From Business Requirement Analysis to MultidimensionalModelling. 1st International Workshop on Requirements Engineering For Business Need And It Alignment, 2005: 44–53.
- [7] J. Schiefer, R. M. Bruckner, B. List. “*A Holistic Approach For Managing Requirements Of Data Warehouse Systems*”. Eight AmericasConference on Information Systems, 2002: 77–87.
- [8] F. Ri. S. Paim, J. F. B. de Castro. *DWARF: An approach for requirements definition and management of Data Warehouse Systems*. Proceedings of the 11th IEEE International Requirements Engineering Conference; 2003: 75–84.
- [9] R. Kimball, M. Ross, W. Thornthwaite, J. Mundy, B. Becker. “The Data Warehouse Lifecycle Toolkit”. Indianapolis: Wiley Publishing, Inc; 2008.
- [10] J.-N. Mazon, J. Trujillo, M. Serrano, M. Piattini. “*Designing Data Warehouses: From Business Requirement Analysis to MultidimensionalModelling*”. 1st International Workshop On Requirements Engineering For Business Need And It Alignment; 2005: 44–53.
- [11] N. Jukic, J. Nicholas. “*A Framework for Requirement Collection and Definition Process for Data Warehousing Projects*”. Proceeding ofthe International Conference on Information Technology Interface; 2010. p. 187–192.
- [12] M. Kumar, A. Gosain, Y. Singh. “Agent Oriented Requirements Engineering for a Data Warehouse ACM SIGSOFT SoftwareEngineering Notes”. *ACM SIGSOFT Software Engineering Notes* 2009; 24(5): 3–6.
- [13] R. Winter, B. Strauch. “*Information requirements engineering for data warehouse systems*”. Proceedings of the 2004 ACM symposium onApplied computing, 2004.
- [14] Shao YiChuan, Xingjia Yao. “Research of Real-time Data Warehouse Storage Strategy Based on Multi-level Caches”. Published in 2012 *International Conference on Solid State Devices and Materials Science* (Physics Procedia 25, 2012: 2315–2321).
- [15] Goguen J. et. al, “*Techniques for Requirements Elicitation*”, Proc. of Int. Symp. on Requirements Engineering, IEEE Computer Society Press, 1993.
- [16] Harrison, M., Zave, P. “*Goal-Driven Requirements Engineering: Modeling and Guidance*” Conference Proceeding of the Second IEEE International Symposium on Requirements Engineering, IEEE Computer Society Press, Los Alamitos, California. 1995 (ed.).
- [17] Haumer P., et. al., “Requirements Elicitation and Validation with Real World Scenes”, *IEEE Transactions on Software Engineering*, 1998; 24(12), Special Issue on Scenario Management.
- [18] Ann M. Hickey, Alan M. Davis “*Elicitation Technique Selection: How Do Experts Do It?*”, IEEE, 2003.
- [19] Eric S. K. Yu “*Towards Modelling and Reasoning Support for Early-Phase Requirements Engineering*”, IEEE, 1997