International Journal of Advances in Computer Science and its Applications – IJCSIA
Volume 4 : Issue 1 *[ISSN 2250 – 3765]*
Publication Date : 09 January 2014

A Review Study on the ASME Criteria for Evaluating Data Integration Approaches

Tg. Aliff Faisal Tg. Ahmad
Department of Computer Science
Faculty of Science & Technology
Universiti Malaysia Terengganu
Kuala Terengganu, Malaysia
taf_2nd@yahoo.co.uk

Mustafa Man
Department of Computer Science
Faculty of Science & Technology
Universiti Malaysia Terengganu
Kuala Terengganu, Malaysia
Mustafaman@umt.edu.my

Md Yazid Mohd Saman
Department of Computer Science
Faculty of Science & Technology
Universiti Malaysia Terengganu
Kuala Terengganu, Malaysia
Yazid@umt.edu.my

***Abstract*— This paper provides an evaluation of some existing data integrations approaches according to the Patrick Ziegler's ASME criteria for evaluating. These criteria for evaluating identify the characteristics of integrated approaches and investigate the existing of data integration approaches for their support for user-specific perspectives. Through this review, this paper aims to identify some potential implementation methods and architectures that support for a truly user-specific perspective in the context of the structured, semistructured, and unstructured data integration approaches.**

***Keywords*— Information integration, Data integration, Data integration approaches, User-Specific Perspectives.**

I. INTRODUCTION

Information integration involves a data integration process that combines data from a heterogenous of data sources in order to provide a unified view of data in a structured form [1][2]. The data integration process is done by applying a global data model and by detecting and resolving schema and data conflicts so that a homogeneous, unified view can be provided [3][4][5].

Information integration involves designing an approach that has the ability to combine data from a variety of independent data sources. Recent years, various specific aspects in information integration approaches [6][7][8] have been studied, proposed and suggested in the pursuit of achieving an ideal data integration system by the research community. Some of these studies claim to resolve the problem. However, most of them meet some difficulties when handling data with less structure and varying granularity [9][10][11]. Information integration, either virtual or materials, provides a mechanism that enables data to be viewed from autonomous, heterogeneous information sources [12][13][14]. From the user perspectives, this mechanism allows the user to query certain number of information sources with the need of knowing neither the distribution of data nor the models and languages exploited for handling them.

The rest of this paper has been structured as follows. Section 2 briefly reviews the ASME criteria that have been defined by Patrick Ziegler. In Section 3, this paper describes four different of existing work and study of data integration. The evaluation results of this paper are presented in Section 4. Finally, the paper concludes in Section 5.

II. THE ASME CRITERIA

According to the Patrick Ziegler's study [3], some of the data integration approaches have been identified in several areas that contribute in his evaluation. Patrick Ziegler defines several areas such as: i) Multidatabase languages and declarative integration languages; ii) Approaches with conceptual-level abstraction from data sources; iii) Object-oriented virtual integration approaches; iv) Ontology-based integration approaches; v) Semantic Web approaches; and vi) Taxonomic database systems This paper only evaluates some approaches from two types of area which are Object oriented virtual integration approaches and Ontology-based integration approaches.

In order to identify and study the characterization of various existing data integration approaches, this paper adapts the ASME technique [3] in the evaluation of data integration approaches. The ASME technique that's been defined by Patrick Ziegler focus to identify user-specific element in data integration approaches. This technique assesses the evaluation each of the existing data integration approaches based on four sets of criteria:

A. Full Abstraction of User from Data Sources

These criteria investigate whether existing work and study of data integration approaches provide a user-friendly interface that can guide and assist users from technical issues of underlying data sources.

B. User-Specific Data Source Selection for Integration

These criteria investigate whether existing work and study of data integration approaches have been designed with some applications that enable users to make data source selection and provide users with the data integration ability from individually for later modeling of tailored views.

C. User-Specific Data Modeling for Integration

These criteria investigate whether the existing data integration approaches contain specified forms. These forms contain some

integration mechanism and enable to fulfill user's information needs. This mechanism also emphasizes the user's way to perceive a domain of interest.

D. Explicit Semantics

These criteria investigate whether existing data integration approaches provide means for explicitly representing the real-world semantics of data.

III. EVALUATIONS OF SELECTED APPROACHES

The characterization of selected integration approaches is shown in Table I. This paper only evaluates four different of existing work and study of data integration; i) The Stanford-IBM Manager of Multiple Information Sources (TSIMMIS); ii) Garlic; iii) COntext INterchange (COIN); and iv) Services and Information Management for decision Systems (SIMS).

TABLE I. THE CHARACTERIZATIONS OF SELECTED APPROACHES

Integration Approaches	Data Representation Types		
	Structured	*Semistructured*	*Unstructured*
TSIMMIS	√	√	√
Garlic	√	√	√
COIN	√	√	
SIMS	√		

From the Object oriented virtual integration approach area, TSIMMIS [15][16][17] at Stanford implements the mediators-based information integration architecture through a simple object exchange model. TSIMMIS has been designed in order to focus on query processing issues pertaining to an integrated system [18]. This approach handles structured, semistructured, and unstructured data integration. Garlic [3] project at IBM Almaden Research Center which targets at developing a system and tools for the management of large quantities of heterogeneous multimedia information. This project is a federated system for diverse data sources. Garlic approach focuses on the design and implementation of heterogeneous multimedia information that involves data from heterogeneous databases, files, multimedia data source types [18][19].

On the other side of Ontology-based integration approach area, COIN [20] approach has been designed to address the problem of semantic interoperability by consolidating distributed data sources and providing a unified view to them. The main focus of COIN design is to provide mediation service modules [5] that integrate data from heterogeneous databases and semistructured web sources [18]. SIMS [21] is an information mediator that provides access and integration of multiple sources of information. The main focus of this approach is to define application domain model using hierarchical terminological knowledge base (Loom) that presents data of structured information source types.

With respect to the ASME criteria, the evaluation of this paper leads to the following result that's been described in Table II.

TABLE II. THE EVALUATIONS OF SELECTED APPROACHES

Integration Approaches	ASME Criteria			
	Abstraction	*Selection*	*Modeling*	*Explicit Semantics*
TSIMMIS	No	Yes	Yes	No
Garlic	No	Yes	Yes	No
COIN	No	Yes	No	No
SIMS	Yes	No	No	No

According to Patrick Ziegler [3], object-oriented virtual integration approaches like TSIMMIS or Garlic have support user-specific that provide some applications for user in data sources selection and modeling. These applications enable users to make individual selection and combine data from data sources. The users receive the information query in model object-oriented.

Garlic approach has been designed with user-friendly interfaces that provide some applications for the user to arrange the display of a subgraph of the object graph formed by the inter-object references. The user's display contains a consistent view of the subgraph of interest. Furthermore, Garlic also been provided with some application that enable the user to indicate which attributes and relationships to be viewed on the current display.

However, the query processing in object-oriented virtual integration approaches has limitation in abstraction criteria. Both of these approaches do not provide an abstract to the user from technical-level heterogeneities. Moreover, users were insufficiently abstract from underlying data sources and thus have to cope with low-level heterogeneities [3][22]. Furthermore both of these object-oriented virtual integration approaches also have lacked support for explicit, queryable semantics of all available data [3].

For example TSIMMIS's constraint management that has been designed to distribute heterogeneous environment's address is more difficult and complex problem than constraint management in centralized systems. The Query transaction description across multiple information sources usually not been provided. Moreover each information source may support different capabilities for assessing and monitoring the data involves in a constraint [5]. Garlic approach does not provide any guarantees about the consistency of legacy data that is operated on by legacy (as well as Garlic) applications. Garlic has a disability in requiring such reference attributes to be converted into and stored by using Garlic's full weak identifier

On the other side of Ontology-based integration approach area, based on ASME criteria, this paper concludes that the design approach from this area seem not aiming at user-specific elements. The result from this area shows that COIN approach only supports user-specific data source selection for integration criteria and SIMS approach only supports in full abstraction of user from data sources criteria. As mentioned previously, a COIN approach only supports in user's data source selection. For example queries in the COIN framework are source-specific. A user formulates a query identifying explicitly the source and attributes referenced.

Based on full abstraction of user from data sources criteria result, from the selected approaches that been evaluated by ASME criteria, SIMS is the only approach that supports full abstraction of user from data sources criteria. As in COIN, this approach does not provide an abstract to the user from technical-level heterogeneities. For example this approach does not scale-up effectively given the complexity involved in constructing a shared schema for a large number of systems and were generally unresponsive to changes for the same reason.

As with SIMS approach, the user is not presumed to know how information is distributed over the data and knowledge bases to which SIMS has access. Hence, the user assumes to be familiar with the application domain, and to use standard terminology in order to compose the Loom query [21].

Based on explicit, queryable semantics criteria result, both of these approaches have lacked support in these criteria. For example COIN contains data semantics that encapsulate in a share schema. These semantics cannot be easily extracted by a user in order to assist in formulating a query which seeks to reference the source schemas [20].

IV. CONCLUSION

This paper provides a brief review and describes some of the existing approaches for the design of data integration. From that, this paper has searched and reviews various research documentation that report related work and present previous study that focus on the models and approaches in the context of data integration.

From reviewing several articles, report, book, and others that focus on the data integration, the finding of this paper record that most designed framework and methods of the existing work are based on the architecture perspectives without concerning the implication to the user-specific. Some of these related work claims to have designed an approach that differs with other approaches and meet some challenge in data integration. Realistically these existing works were just making an improvement features from the previous work and aiming to find solution of issues and challenges that been described in the previous work. Most of the study that review of previous research report just describing the general problem of information integration and discussing the challenges that need to be addressed in order to deal with the general problem of information retrieval and integration.

The methods that proposed by Patrick Ziegler show the different way in evaluating existing work and study. The ASME criteria not just identifying the characteristics of existing approaches, but also investigates existing of data integration approaches for their support for user-specific.

This paper evaluates existing data integration approaches for their compliance with the ASME criteria (Abstraction, Selection, Modeling, and Explicit semantics). In retrospect to the findings of the evaluation, from Table 1 it can be seen that none of the presented approaches are able to fully attain the ASME criteria especially in explicit, queryable semantics criteria which the result shows none of the presented approaches provides user with explicit statements that represent the real-world semantics of data. Without explicit statements on the intend semantics, users will find some difficulties in interpreting data and schema items themselves. Due to this issue, misinterpretations result will occur and some important underlying assumptions concerning source data may be fully implicit.

ACKNOWLEDGMENT

I would like to thank Dr. Mustafa Man and Professor Md Yazid Mohd Saman from Faculty of Science & Technology, Universiti Malaysia Terengganu for many deep and valuable discussions and also constructive guidance in the completion of this paper.

REFERENCES

[1] M. Friedman, A. Y. Levy, and T. D. Millstein, "Navigational Plans For Data Integration." In *AAAI/IAAI*, 1999, pp. 67–73.

[2] M. R. Genesereth, A. M. Keller, and O. M. Duschka, "Infomaster: An Information Integration System." in *SIGMOD Conference*, 1997, pp. 539–542.

[3] Ziegler, Patrick " User-Specific Semantic Integration of Heterogeneous Data: What Remains to be Done? " *Building the Information Society* 156 (2004): 3-12. Print.

[4] M. Lenzerini, "Data Integration: A Theoretical Perspective." in *PODS, 2002, pp. 233–246.*

[5] Zachary G. Ives, Daniela Florescu, Marc Friedman, Alon Levy, and Daniel S. Weld. 1999. An adaptive query execution system for data integration. *SIGMOD Rec.* 28, 2 (June 1999), 299-310. DOI=10.1145/304181.304209 http://doi.acm.org/10.1145/304181.304209

[6] Jayaraman, Gayathri. A mediator-based data integration system for query answering using an optimized extended inverse rules algorithm. Ontario,: Ottawa,, 2010. Print.

[7] K. C.-C. Chang, B. He, and Z. Zhang, "MetaQuerier over the Deep Web: Shallow Integration across Holistic Sources." in *IIWeb - VLDB*, 2004.

[8] Sonia Bergamaschi, Francesco Guerra, and Maurizio Vincini. A Peer-to-Peer Information System for the Semantic Web. In Proceedings of the *International Workshop on Agents and Peer-to-Peer Computing (AP2PC 2003)*, July 2003.

[9] B. He, M. Patel, C.-C. Chang, and Z. Zhang, *"Accessing the Deep Web: A Survey."* University of Illinois, Urbana-Champaign, Tech. Rep., 2004.

[10] B. He, Z. Zhang, and K. C.-C. Chang, "Towards Building a MetaQuerier: Extracting and Matching Web Query Interfaces." in *ICDE*, 2005, pp. 1098–1099.

[11] D. Florescu, A. Levy, and A. Mendelzon, *"Database techniques for the world-wide web: A survey."* SIGMOD Record, vol. 27, no. 3, pp. 59–74, 1998.

[12] G. Kabra, C. Li, and K. C.-C. Chang, "Query Routing: Finding Ways in the Maze of the Deep Web." in *International Workshop on Challenges in Web Information Retrieval and Integration*, 2005, pp. 64–73.

[13] J. Reinoso, A. Silvescu, D. Caragea, J. Pathak, and V. Honavar, "Information Extraction and Integration from Heterogeneous, Distributed, Autonomous Information Sources: A Federated Ontology-Driven Query-Centric Approach." in *IRI*, 2003, pp. 183–191.

[14] Koch, Christoph. Data integration against multiple evolving autonomous schemata. Meyrin: CERN, 2001. Print.

[15] Chawathe S., H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ull-man, and J. Widom [1994]. "The TSIMMIS project: integration of heterogeneous information sources," *IPSJ Conference*, Tokyo, 1994. Available by anonymous ftp as pub/chawathe/1994/tsimmis-overview.ps from db.stanford.edu.

[16] Papakonstantinou, Ioannis G. "Query processing in heterogeneous information sources". PhD Thesis, Stanfort University, 1997.

[17] S. Kambhampati, U. Nambiar, Z. Nie, and S. Vaddi, *"Havasu: A Multi-Objective, Adaptive Query Processing Framework for Web Data Integration."* Arizona State University, Tech. Rep., 2002.

[18] A. Telang and S. Chakravarthy, "Information Integration across Heterogeneous Domains: Current Scenario, Challenges and The InfoMosaic Approach" in *First International Workshop on Ranking in Databases in Conjunction with ICDE 2007*, 2007.

[19] Ruxandra Domenig and Klaus R. Dittrich. 2000. A query based approach for integrating heterogeneous data sources. In Proceedings of the ninth international conference on Information and knowledge management (CIKM '00). ACM, New York, NY, USA, 453-460. DOI=10.1145/354756.354853 http://doi.acm.org/10.1145/354756.354853

[20] Goh, Cheng, St Bressan, Stuart Madnick, and Michael Siegel. "Context interchange: new features and formalisms for the intelligent integration of information." *ACM Trans. Inf. Syst.* 17.3 (1999): 270-293. Print.

[21] Yigal Arens and Craig Knoblock. 1993. SIMS: Retrieving and integrating information from multiple sources. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data* (SIGMOD '93), Peter Buneman and Sushil Jajodia (Eds.). ACM, New York, NY, USA, 562-563. DOI=10.1145/170035.171566 http://doi.acm.org/10.1145/170035.171566

[22] Xiao, Huiyong. Query processing for heterogeneous data integration using ontologies. Chicago: Illinois, 2006. Print.