RESEARCH ARTICLE OPEN ACCESS

Study on Theoretical Aspects of Virtual Data Integration and its Applications

Mrs. Munmun Bhattacharya*, Nashreen Nesa**
*(Department of Information Technology, Jadavpur University, Kolkata)
** (Department of Information Technology, Jadavpur University, Kolkata)

ABSTRACT
Data integration is the technique of merging data residing at different sources at different locations, and providing users with an integrated, reconciled view of these data. Such unified view is called global or mediated schema. It represents the intentional level of the integrated and reconciled data. In the data integration system, our area of interest in this paper is characterized by an architecture based on a global schema and a set of sources or source schemas. The objective of this paper is to provide a study on the theoretical aspects of data integration systems and to present a comprehensive review of the applications of data integration in various fields including biomedicine, environment, and social networks. It also discusses a privacy framework for protecting user's privacy with privacy views and privacy policies.
Keywords - Data integration, view, schema, biomedicine, social, privacy

1. INTRODUCTION

Over the past few years, there have been significant changes in the role of the database and especially of database techniques. Today's computer networks provide us with access to a wide variety of databases. In some cases, the answer to our queries is contained in a single database. But in the majority of cases the necessary data, as an answer to our query is distributed over several sources. Thus, it becomes indispensible to merge data from these multiple sources and fulfill the user's request, resulting in the need for a Data Integration process. Data integration is the technique of combining data residing at different sources, and providing the user with a unified view of these scattered data. Such unified view is called global schema. It represents the intentional level of the integrated and reconciled data, and provides the means for expressing user queries. In formulating the queries, the user is freed from all the internal knowledge of the necessary data. For example, where the data resides, how the data are structured at the sources, and how the data are to be merged and reconciled. Designing data integration systems is important in current real world applications, and is characterized by a number of issues that are interesting from a theoretical point of view. The area of interest in this paper is characterized by an architecture based on a global schema and a set of sources. The sources contain the real data, while the global schema provides a reconciled, integrated, and virtual view of the underlying sources. Modeling the relation between the sources and the global schema is therefore a crucial aspect and requires an in-depth knowledge of each component. Two basic schemas have been proposed to this purpose. The first schema, called global-as-view, requires that the global schema is expressed in terms of the data sources. The second, called local-as-view requires the global schema to be specified independently from the sources, and the relationships between the global schema and the sources are established by defining every source as a view over the global schema. Irrespectively of the method used for the specification of the mapping between the global schema and the sources, one basic service provided by the data integration system is to answer queries posed in terms of the global schema. Given the architecture of the system, query processing in data integration requires a reformulation step: the query over the global schema has to be reformulated in terms of a set of queries over the sources. In this paper, such a reformulation problem will be analyzed for both the case of local-as-view, and the case of global-as-view mappings. In the ideal case, users would want to pose queries on a single data integration system and have the system automatically configure itself to integrate data from a set of data sources so that it can correctly and efficiently answer queries that span multiple sources. In order to achieve this, efficient tools should be built to reduce the effort required to integrate data sources. For example, these tools should make it easy to add a new data source, relate its schema to others, and automatically adjust the data integration system for better performance. Another way to achieve it is to improve the ability of the system to answer queries in uncertain environments. In applications where data integration facilitates exploratory efforts (e.g., on the Web), the system should be able to answer queries under uncertainty. For example, when we look for an

article on the Web, it's okay to find 9 out of 10 relevant sources, or to return answers that don't completely satisfy the user query.

2. DATA INTEGRATION ARCHITECTURE

In order to fully understand how a Data Integration System works, it is important to understand the components which make it. The architecture of a Virtual Data Integration system includes the following components. [1]

Data Sources- Data sources can vary on many dimensions, such as the data model underlying them and the kinds of queries they support. Examples of structured sources include database systems with SQL capabilities, XML databases with an XQuery interface, and sources behind Web forms that support a limited set of queries. In some cases, the source can be an actual application that is driven by a database, such as an accounting system. In such a case, a query to the data source may actually involve an application processing some data stored in the source.

Mediated (Global) Schema- The user interacts with the data integration system through a single schema, called the mediated schema. The mediated schema is built for the data integration application and contains only the aspects of the domain that are relevant to the application. That is, it does not necessarily contain all the attributes we see in the sources, but only a subset of them. In the virtual approach, the mediated schema is not meant to store any data at all. It is purely a logical schema that is used for posing queries by the users employing the data integration system.

Source Description- The key to building a data integration application is the source descriptions that connect the mediated schema and the schemas of the sources. These descriptions specify the properties of the sources that the system needs to know in order to use their data.

Semantic Mappings- The semantic mappings are the main components of source descriptions, which relate the schemata of the data sources to the mediated schema. The semantic mappings specify how attributes in the sources correspond to attributes in the mediated schema (when such correspondences exist).

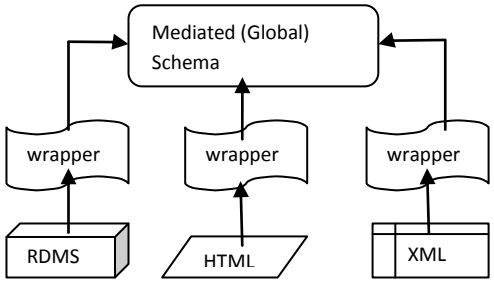

Fig 1: Data Integration components

Thus, a logical framework can now be set up for Data Integration. We restrict our attention to data integration systems based on a so-called global schema (or, mediated schema). In other words, we refer to data integration systems whose aim is combining the data residing at different sources, and providing the user with a unified view of these data. Such a unified view is represented by the global schema, and provides a reconciled view of all data, which can be queried by the user. Fig 1 illustrates all the components in a typical virtual data integration system. The main task in the design of a data integration system is to establish the mapping between the data sources and the global schema.

3. MODELLING OF DATA INTEGRATION SYSTEMS

One of the most important aspects in the design of a data integration system is the specification of the correspondence between the data sources and those in the global schema. Such a correspondence is modeled through the notion of mapping as introduced in the previous section. It is exactly this correspondence that will determine how the queries posed to the system are answered. Two basic approaches for specifying the mapping in a data integration system have been discussed, called local-as-view (LAV), and global-as-view (GAV). Assuming S_1,..., S_n to be the local schemas of n pre-existing data sources and G_1,...,G_m to be m global relations of the global schema G. The aim is to model semantic relations between the local schema S_i and the global schema G_j [2].

3.1 Local-As-View (LAV)

The semantic mappings are of the form

$$S_i \subseteq V_i (G_{1, \ldots\ldots, } G_m)$$

where each Vi is a view over the global schema, i.e., a query built on global relations.[2]

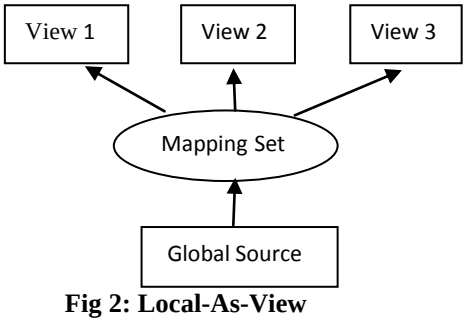

Fig 2: Local-As-View

In this approach, the source schemas are modeled as a set of views over an underlying global schema. The local-as-view method, does not describe the global schema in a direct manner but in a way which is disassociated from the data sources. The mapping between the global schema and the data

sources is created by creating views of each source over the global schema. The advantage of this model is that new sources can be added easily when compared to GAV. However the query rewriting process is complex because the system has to choose from a set of choices to determine the best possible rewrite.

3.2 Global-As-View (GAV)

The semantic mappings are of the form

$$V_i(S_1,\ldots\ldots,S_n) \subseteq G_i$$

also equivalently denoted as

$$G_i \supseteq V_i(S_1 \ldots\ldots,S_n)$$

where each Vi is a view over the local schemas, i.e. a query built on local relations[2]. In this approach, the global schema is modeled as a set of views over the source database. The advantage of this model is that the query processing is simpler. But the disadvantage is that the addition of a new source needs considerable effort. This makes GAV a good choice when the sources are less probable to change.

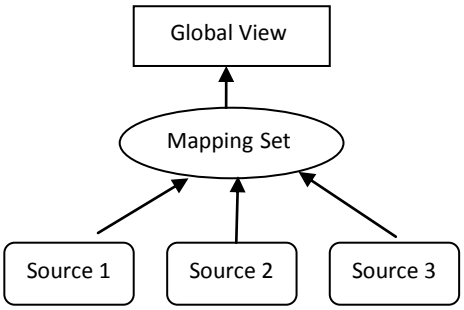

Fig 3: Global-As-View

3.3 Comparison between LAV and GAV

In both the GAV and LAV approach it can be observed that the basic characteristic of being able to answer queries in terms of the global schema exists. Furthermore, in both cases queries over the mediated schema have to be translated in terms of a set of queries over the component data sources. However, query processing using the local-as-view approach is difficult due to the fact that information about the data in the global schema is gained through views that correspond to the component databases, thus we have fragmented information about the data. Querying in the GAV approach is significantly simplified, this is because we are easily able to understand which source queries map to the mediated schema. As previously mentioned, adding new data sources in the LAV approach is easy. This is due to the fact that describing new sources is not dependent on knowing anything about other sources and the associations that exist between these sources. In GAV, adding another component database is difficult because views need to be updated. With LAV, creating more definitive descriptions of the data sources is possible.[3]

4. PRIVACY FRAMEWORK FOR DATA INTEGRATION

Database security has generally focused on access control. Users are explicitly (or implicitly) allowed certain privileges to a data item. Clifton et al. proposed a framework [4] defining the private data and privacy policies in the context of data integration and sharing. The notion of Privacy Views and Privacy Policies is essential towards such a framework.

4.1 Privacy Views

The database administrator specifies a set of privacy views that defines what is private data in a declarative language extending SQL. Each privacy view specifies a set of private attributes and an owner. According to this definition, data that appears in the privacy view is considered private; otherwise it is not private. For example, the database administrator in an educational organization might define the following privacy view:

PRIVACY-VIEW studentAddressDob
OWNER student.uid
SELECT Student. address, Student.dob
FROM Student

The above query specifies that a student's address and dob (date-of-birth) are considered private data when occurring together. Whenever these two attributes occur together in a piece of data, e.g., to integrate with other student's data, they are private. It is to be noted here dob is not private by itself (and similarly address). Privacy views could be implemented by a privacy monitor that checks every data item being retrieved from the database and detects if it contains items that have been defined as private.

4.2 Privacy Policies

In addition to privacy views privacy, a notion of privacy policies is also implemented. The privacy policies are decided by the database administrator for each view. Continuing with the example, the following privacy policies could be specified:

PRIVACY-POLICY studentData
ALLOW-ACCESS-TO y
FROM Consent x, studentAddressDob y
WHERE x.pid = y.owner and x.type = 'yes'
BENEFICIARY *

The above privacy policy states that private data studentAddressDob can be released if the owner has given explicit consent, as registered in a Consent table.

5. RELATED WORKS

5.1 Data Integration in Biomedicine

We are experiencing the emergence of the "data rich" era in biology [5]. The explosion in the number and size of biomedical data, and the rapid growth in the variety and volume of laboratory data has been fuelled by world-wide research activity and the emergence of new technologies. In fact, according to Peter Buneman[5] the data in some biological data sets surpass that of "big science" data by orders of magnitude. Thus the analysis of this data often requires an efficient integration of these heterogeneous and typically unstructured data, distributed across many data sources. One such model proposed by P.Mork et al. [6] is used for efficient integration of biomedical data that is applied to online genetic databases. By successfully creating a mediated schema, which is a graphical representation of the entities in the domains, this model effectively returns the gene/protein pairs that are believed to be associated to a disease given in the query string. Similar work proposed by Kang et al. is based on integration of heterogeneous microarray gene expression datasets [5]. Here, the gene expression data are organized as tables. The proposed algorithm captures the system-wide dependencies existing between the genes in a microarray data across disparate sources, and then compares the signatures across the tables for further analysis. [7, 8, 9,10] basically deals with integration of biological data but varies on several dimensions. For eg, Greeshma et at. assigns unique keys to chemical compounds using SMILES (Simplified Molecular Input Line Entry System) notation for integrating data concerning chemical structures. These identifiers can then be used as database keys [7]. Seong Joon et al. implemented SOAP API for integrating biological interaction databases [9]. Astakhov et al. developed a Biomedical Informatics Research Network (BIRN) project to create a multi-institution information management system that integrates data from each participating institutions and performs analyses that could not be executed from any single institution's data [10]. In BioLog: a Browser Based Collaboration and Resource Navigation Assistant, P. Singh et al. describe a platform for biomedical researchers that extracts access patterns of researchers as they browse PubMed, a repository for peer-reviewed research reports in the field of life sciences. BioLog makes recommendations that use a combination of collaborative filtering and content based filtering techniques [5]. Cohen et al. developed BioGuide, an information retrieval system, which helps scientists to choose suitable sources and tools, find complementary information in sources, and deal with divergent data [5].

5.2 Data Integration in Social Networks

The accessibility of huge amounts of personal data in a social network is an excellent source for studying human behaviour and interactions. For e.g., bookmarking and tagging data suggests user interests, frequency of commenting suggests the strength of relations, etc. In order to perform data mining to exploit these data to its fullest potential, a unified view of such data from different social networks is required. This is exactly the place where data integration of social networks becomes essential. Tang et al. in their work [11] present a joint optimization framework to integrate multiple data sources for community detection. This work is inspired by the fact that groups in a social network share more similarities or interact more frequently among themselves than with people outside groups. Unfortunately, privacy and security are major concerns as the data published over the web still remain open and there are no means for monitoring any unauthorized access to data in social networks. This issue motivated researchers to include privacy as a main feature in data integration. Barth et al. proposed a formal framework for communicating privacy expectations and privacy practices, inspired by contextual integrity. Contextual integrity is a term used for an account of privacy in terms of the transfer of personal information [12]. Somewhat along the same line, Kayes et al. developed Aegis [13], a model incorporating privacy protection as contextual integrity by using semantic web tools. In addition, it defines default privacy policies that protect a user's private information from other users. This work is motivated by the fact that users usually retain their default privacy settings even though they are invited to change, but in reality very few do it. Sun et al. work involves security in data sharing in social networks [14]. In this paper, they propose a privacy-preserving scheme for this purpose with efficient measures for preventing a user's access right to any private data once the user is removed from their social group. In Frientegrity, a framework for social networking application, Feldman et al. attempted to build a secured application framework that deals with an untrusted service provider. In light of the threat that these service providers are entrusted with the privacy of one's social interactions and they may not always be deserving of that trust. Here, a novel method is presented for detecting server equivocation in which users themselves collaborate to validate object histories [15].

5.3 Data Integration in Environmental scenarios

Tran et al. presented their work in the scope of a project named ADMIRE (Advanced Data Mining and Integration Research for Europe) [16]. This project proposes a framework for integration and mining of environmental data. ADMIRE is motivated by the

difficulty in the extraction of meaningful information related to environmental scenarios from multiple heterogeneous and distributed sources. The project has been well tested in the specific scenarios that are part of the Flood Forecasting and Simulation Cascade application. Other works include that of Deng et al. that provides a model called Dynamics of Land Systems (DLS) that is capable of integrating multiple data sources to simulate the dynamics of a land system [17]. Michener et al. designed SEEK (Science Environment for Ecological Knowledge), a project to help ecologists overcome data integration and synthesis challenges. The SEEK is capable of capturing, organizing, and searching data for complex scientific analyses [5].

5.4 Data Integration for Business Intelligence

Business Intelligence allows an organization to gain a better understanding of their clients, the market, supply, as well as competitors in order to make smart business decisions. Chung et al. in their work [18] designed tools for integrating and mining a company's web site in order to gain crucial strategic decisions. Two case studies have been discussed for this purpose; the first study is done using relational database techniques such as Oracle database and Cognos BI tool. The second case is for multimedia data analytics using Monago database and Pentaho BI tool for integrating and mining multimedia data presented for the travel related analytics of Food & Wine website. J.Madhavan et al. proposed a web-scale based data integration system [19] named PayGO that incorporates new techniques to handle the scale and heterogeneity of structured web data. The pay-as-you-go principle states that the system needs to be able to incrementally evolve its understanding of the data it encompasses as it runs. Here, instead of a single mediated schema, repository of schemata is present for answering queries that are clustered by topic. This work is inspired by the concept of data spaces and emphasizes pay-as-you-go data management as means for achieving web-scale data integration.

6. CONCLUSION

A major issue today is that crucial information is scattered throughout the separately developed data sources, in a way that makes the "big picture" difficult to obtain. Data integration presents a unified virtual view of all these scattered data within a domain, allowing users to pose queries across the complete integrated schema as if they are interacting with a single data source. In this paper we have highlighted some of the theoretical issues underlying data integration. This study identifies the two main modelling techniques of query reformulation in data integration system model that is global-As-view and local-As-view. We examined the definitions and architecture of each component in the data integration system. The study also discusses a privacy framework for data integration that was proposed by Clifton et al. [4]. In addition, we have presented a comprehensive review of real-world applications of virtual data integration in various fields of biomedicine, social networks, environment and business intelligence. The discussion covers the works and applications of data integration, and why data integration is a necessity today. Although many aspects of data integration have been studied in different disciplines, the study of data integration in environment and ecology is still in an early stage.

REFERENCES

[1] AnHai Doan, Alon Y. Halevy, Zachary G. Ives: *Principles of Data Integration* (Morgan Kaufmann 2012).

[2] Abiteboul, Serge, et al. *Web data management*. (Cambridge University Press, 2011).

[3] Tsierkezos, *S. Comparing Data Integration Algorithms Dis Thesis abstract.* University of Manchester 2014.

[4] C.Clifton, M.Kantarcioğlu, A. Doan, G. Schadow, J.Vaidya, A.Elmagarmid, and D.Suciu, "Privacy preserving data integration and sharing," in *Proceedings of the 9th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery* - DMKD '04, 2004, p. 19.

[5] Baker, Christopher JO, Gregory Butler, and Igor Jurisica, eds. *Data Integration in the Life Sciences: 9th International Conference, DILS 2013, Montreal, Canada, July 11-12, 2013, Proceedings*. Vol. 7970. Springer, 2013.

[6] P.Mork ,A.Halevy, P.Tarczy-Homoch *"*A Model for Data Integration Systems of Biomedical Data Applied to Online Genetic Databases*"* in *Journal of the American Medical Informatics Association*, Fall Symposium Supplement(2001). p.473-477

[7] Greeshma Neglur and Robert L. Grossman, "Assigning Unique Keys to Chemical Compounds for Data Integration: Some Interesting Counter Examples", *2nd International Workshop on Data Integration in the Life Sciences* (DILS 2005), La Jolla, July 20-22, 2005.

[8] T. Kirsten, H.-H. Do, C. Körner, E. Rahm: "Hybrid integration of molecular biological" annotation data. *Proc. Intl. Workshop on Data Integration in the Life Sciences,* 2005.

[9] Yoo, Seong Joon, et al. "SOAP API for integrating biological interaction

databases." *Data Integration in the Life Sciences. Springer Berlin Heidelberg*, 2005. pp 305-308

[10] Astakhov, V., Gupta, A., Santini, S., & Grethe, J.S.(2005, January). "Data integration in the biomedical informatics research network(BIRN")*.*In *Data Integration in the Life Sciences* (pp. 317-320). Springer Berlin Heidelberg.

[11] Tang, Jiliang, Xufei Wang, and Huan Liu. "Integrating social media data for community detection." *Modeling and Mining Ubiquitous Social Media.* Springer Berlin Heidelberg, 2012. 1-20.

[12] Barth, A., Datta, A., Mitchell, J. C., & Nissenbaum, H. (2006, May)."Privacy and contextual integrity: Framework and applications."*In Security and Privacy, 2006 IEEE Symposium* (pp. 15-pp). IEEE.

[13] Kayes, Imrul, and Adriana Iamnitchi. "Aegis: A semantic implementation of privacy as contextual integrity in social ecosystems." *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on*. IEEE, 2013.

[14] J. Sun, X. Zhu, and Y. Fang. "A privacy-preserving scheme for online social networks with efficient revocation." *In Proc. INFOCOM, Mar. 2010.*

[15] Feldman, A. J., Blankstein, A., Freedman, M. J., & Felten, E. W. (2012, August). "Social Networking with Frientegrity: Privacy and Integrity with an Untrusted Provider." *In USENIX Security Symposium* (pp. 647-662).

[16] Tran V.D., Hluchý L., Habala O. "Data mining and integration for environmental scenarios" *Proc. 2010 Symp. on Information and Communication Technology*, Hanoi, Vietnam, 27–28 August 2010 55 58 New York, NY ACM

[17] Deng, Xiangzheng, Hongbo Su, and Jinyan Zhan. "Integration of multiple data sources to simulate the dynamics of land systems." *Sensors* 8.2 (2008): 620-634.

[18] Ping-Tsai Chung, Sarah H. Chung "Data Integration and Data Mining for Developing Business Intelligence" *Farmingdale, NY 3-3 May 2013 pages 1-6*

[19] Jayant Madhavan, Shawn R.Jeffery, Shirley Cohen, Xin (Luna) Dong, David Ko, Cong Yu, Alon Halevy "*Web-scale Data Integration: You can only afford to Pay As You Go*" CIDR (2007)