

# Data Representation, Extraction and Integration Process For web and Database

Muhammad Shoaib, Shazia Arshad and Prof. Dr. Abad Shah  
Department of Computer Science and Engineering  
University of Engineering and Technology,  
Lahore, Pakistan

## Abstract

*The WWW is considered as a collection of heterogeneous information sources available online. However this information (web sources) need to be integrated and inquired uniformly, and transparently from direct intercalation with web sources detailed. This problem called a data integration problem, where uniform access to a multiple heterogeneous information is desired. In fact, this problem has received a considered attention from researcher in field of database systems. In this report we discuss the Information integration system approach that is used in many existing integrating tools and we propose a new hybrid approach to overcome the drawbacks of query optimization found in that approaches. Also a proposed solution to an outstanding problem in the information integration process is discussed with a solution solved some of these problem.*

**Keywords:** Data integration, Source description, Wrapper, Mediated Schema, Local Schema, GDM

## 1. Introduction

An increasing amount of data is becoming available online to the web users, where the data is managed using different data models and access mechanisms.

In fact, the WWW contain a number of information sources (web sources that can be viewed as a containers of sets of tuples. However, the task of information integration process is to answer user queries uniformly by integrating multiple web sources. But this task is faced with fact that the web sources is considered as a semistructured data which is already stored in sources

that do not have structure or when data is combined from several heterogeneous data sources.

Most research of building information integration system is considered one of the two approaches; the first approach is data warehouse, where the information is integrated in a single repository for querying later on. And the change in the data may not be visible online. The second approach is virtual approach, where the information is retrieved on demand (when the user initiated a query).

In the context of information integration and extracting process, many outstanding problems can be summarized as follow:

- 1) In the data integration system the need for creation source description is essential which is used in the purpose of query reformulation. And there is a problem on the design a framework on creation of the source description process. And also in the creation process itself.
- 2) For the reason of no or bad metadata and no statistical data to a query optimizer to generate query execution plans. Thus no good query execution plan can be generated. And poor query optimization is need to be improved even when unexpected situations are happened.
- 3) In the Information integration process the handling of violation of the schema is not discussed.
- 4) In the context of extracting data or tuples from web sources, the current research is emphasizing on the syntactical structure of the web pages rather than on semantics.
- 5) The problem of schema matching between mediated schema and source schema (Local schema).
- 6) In constructing the wrapper to extract the data from the web resource, a problem of invisible data within the

HTML pages where it is designed for human viewing rather than extracting tools. In this report we propose hybrid information integration system architecture by applying a hybrid approach to overcome some of the drawbacks of the other approaches. And we also solve and demonstrate of how to apply schema matching mechanisms to translate from mediated schema to source schema (Local schema) which is considered as a one of the outstanding problem in the scope of information integration process.

The structured of the report is as follows: we begin in Section 2 by discussing the related work and background in the context of information extraction and integration process for the web domain. Also a comparison between information integration system and both, distributed DB and traditional DB is summarized. However, a typical architecture of virtual data integration system is discussed with brief description of its main functions. At the section end a modeling of underlying heterogeneous domain like web sources is argued in both two modeling process (Semi-structured and Graph Data Model). In Section 3 we propose a new approach we define it as hybrid approach in the context of building a data integration system. Section 4 present the problem associated with of the functions of information integrated system, where it is to apply the mapping between mediated schema and source schema, this problem is called schema matching problem. In this section a presentation with proposed mechanisms with demonstrated example for automated schema matching process is done. Finally, Section 5 concludes the work with perspectives and directions for future research.

## 2.Related Work And Background:

Many tools are implemented for the purpose of integrating heterogeneous information sources founds in the web sources. However, an implemented tools called IM (Information Manifold), which considered the WWW as a structured information sources (e.g. Database... etc) accessed within specified form. Where in fact it is a collection of interconnected collection of unstructured documents. And this tools implemented to provide a uniform interface to the structured information on the web [1].

In [2] a design and implementation of integrated system of heterogeneous information sources is presented, where a high level SQL like called W3QL language is used to support query processing.

An another attempt of information querying across a heterogeneous information is at [3]. Which develop a DIOM (Distributed Interoperable Object Mode) framework for the purpose of building an adaptive query mediation framework.

However, these above related work are considering one of the two approaches: warehousing or virtual approach. Where in this report we combine the two approach which is called hybrid approach to overcome the drawbacks of previous approaches.

### 2.1 Data Integration system Vs Distributed DB and Traditional DB:

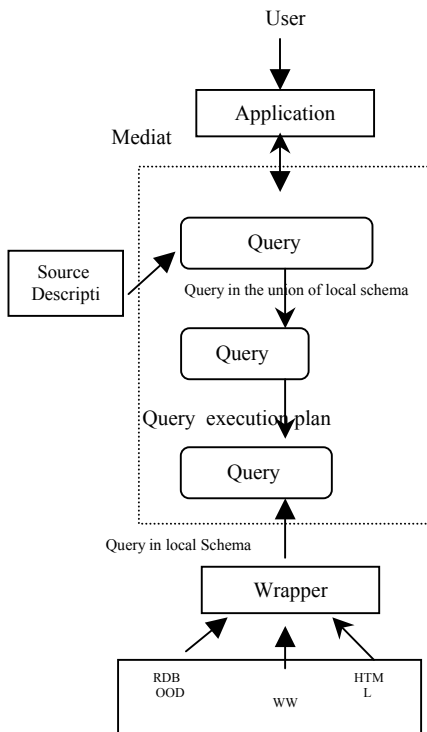
Summarization of the differences between the traditional database context and that of data integration system are illustrated in (Table 1)

Traditional Database	Data integration system
Communicate directly with resource management (Local Storage manger) in order to obtain data..	The query execution engine communicate through Data Integration System.
The user pose the query with associated schema ( <i>Static Schema</i> ).	The user dos not pose queries directly in the schema in witch the data is stored, Instead the query is posed in an <i>mediated Schema</i> (Dynamic schema).
There is An integrity constraints.	No Integrity constraints.
Uniform Structure.	No Uniform Structure..
Standard query language and data model	No standard query language and data model.
Statistical about the source is incrementally and the optimizer use this information to compare between different plans.	Since the source are autonomous The optimizer doesn't have statistics about the source .
The optimizer can reliably estimate the time to transfer data from the disk to main memory.	Data is often transferred over wide area network, and delays may occur for multitude of reasons.

**Table 1:** Differences between the traditional database and data integration system.

## 2.2 Architecture of virtual data integration system :

Architecture of a virtual data integration system is



shown in Figure 1.

**Figure 1:** Architecture of virtual Data integration.

A user of data integration system initiate queries in terms of the mediated schema. And the data integration system uses the source descriptions in order to reformulate a user query into a query that refers directly to the schemas of the sources. However the traditional query execution engine is communicate with a local storage manager to manipulate the data, where in the data integration system the data is manipulated from external resources. So in this case, a wrapper is used which is dedicated for a specific web sources [5], [6], [7]. And the task of the wrapper is to translate the data from the web source to the form usable by the query processor. Which is done by applying the query to the web source and extract a tuples out of the resulting HTML pages.

In general, The most important advantage of a data integration system is that it enables users to focus on specifying what they want, rather than thinking about how to obtain the answers. As a result, the users are working transparently from the data sources and interaction is within a particular interface.

## 2.3 Modeling and Querying the Web/DB:

In fact we need to model the underlying domain such (web itself, structure of web sites, internal structure of web pages, and finally, contents of web sites). In this section we discuss the main distinguishing factors of the data models used in web applications.

**Semi-structured data models:** The first aspect of modeling data for web applications is that in many cases the structure of the data is irregular. Specifically, when modeling the structure of a web site, we don't have a fixed schema, which is given in advance. When modeling data coming from multiple sources, the representation of some attributes (e.g. addresses) may differ from source to source [4], [8], [9].

**Graph data models:** this model is widely in many applications, where it is require to model the set of web pages and the links between them. These pages can either be on several sites or within a single site. Hence, a natural way to model this data is based on a labeled graph data model. Specifically, in this model, nodes represent web pages (or internal components of web pages), and arcs represent links between pages. The labels on the arcs can be viewed as attribute names. One central feature that is common to these query languages is the ability to formulate regular path expression queries over the graph. Regular path expressions enable posing navigational queries over the graph structure.

However, any proposed new model should include web itself, structure of web sites, internal structure of web pages, and contents of web sites. One commonly used model is graph data model. Specifically, in this model nodes represent web, and arcs represent links between them. The labels on the arcs can be viewed as attribute names. Several query languages that is used graph data model have been developed, such as WebSQL, W3QL and WebLog [2].

Suppose a database is represented as directed graph, whose set of nodes consists of set of oids and set of value. Labels are attached to edges, where any two nodes  $x, y$  and any label  $a$  there can be at most one edge between  $x$  and  $y$  labeled  $a$ . and we write  $x \rightarrow a \rightarrow y$ . In addition to the graph, the database also contains a number of collections.

Each collection is a set of nodes. And the collection is the entry point to the graph.

Suppose the following data graph (see Figure 2):

**Figure 2:** Person Collection Data graph

Person = {Abdulaziz, Ahmad}		columns ( X Y L Z )		
Abdulaziz	p1	Title	"..."	
Abdulaziz	p1	Abstract	"..."	
Abdulaziz	p2	Title	"..."	
Abdulaziz	p2	Date	9/3/88	
Ahmad	p2	Title	"..."	
Ahman	p2	Date	11/10/200	

And we apply the following query:

The *where* clause produces all tuples  $(X, Y, Z, L)$ , such that  $X$  is a member of the *Person* collection, there exists an arc labeled "*Paper*" or "*Publication*" from  $X$  to  $Y$ , and there is an arc from  $Y$  to  $Z$ . The result of applying the query to this graph is also given. Which returns all papers this data graph, where collection contains the identifier Abdulaziz and Ahmad (see Figure 3)

In fact, A graph data model is appropriate because site data may be derived from multiple sources, such as existing database systems and HTML files. And the graph Data model can be used as global Data model, where in every level in the architecture is viewed uniformly as a graph.

### 3. Hybrid Approach in data integration system:

As we mentioned that the most research on querying heterogeneous information sources using one of the two approaches; a warehousing or a virtual approach. In the warehousing approach, data from multiple web sources is loaded into a warehouse, and all queries are applied to the warehoused data; this requires that the warehouse be updated when data changes, but the advantage is the good processing time.

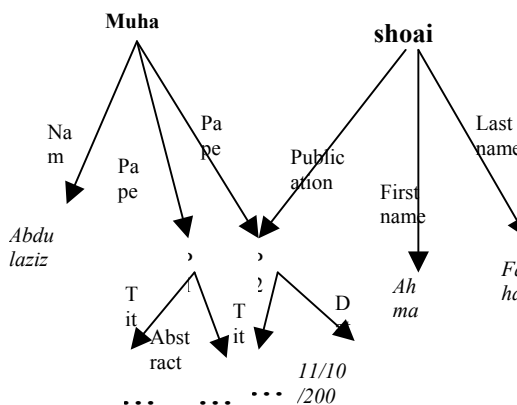
In the virtual approach, the data remains in the web sources, and queries to the data integration system are decomposed at run time into queries on the web sources. And the optimization process is not efficient.

Here in this report we propose hybrid information integration system architecture used a hybrid approach to querying the heterogeneous information sources lie the WWW web sites. In fact, some information in the web sources is stable while the other is changing dynamically. For example consider a database consisting of information about cities. Most geographic information for the city is stable, while other information like weather data is dynamically changed. In this case a hybrid approach is applied, where a data warehousing approach would require weather data to be stored in a local database

**Figure 4:** Architecture of Hybrid Data integration.

One of the functions of *Coordinator* is to every time its data changed. In contrast, the virtual approach will require that all data including stable data to be extracted from external sources.

In the proposed hybrid approach it allow the stable information to be stored permanently in a local database, while the dynamic information is extracted using virtual approach method.



Class	oid	value
Artyicle	P	tuple ( title : "test pasge" Author : list ( AS, SA, FM))
Author	AS	A.solieman
Author	SA	S. Ahmad
Author	FM	F.Mohamad
Section	S!	tuple ( section 1 : tuple ( title : "Introduction ....",

The architecture of the Hybrid Integration System Is illustrated in Figure 4 bellow:

perform the integration between the information stored in the external sources and a local database. Another is to construct a requests to the *load utility* to embedded a new external sources in the *data repository*, where the *Load Utility* in his term initiated the *wrapper* to connect with the external resource and apply the result back to the utility loader. In this case the embedding stage is done during the query processing. One benefit from hybrid approach is to solve the problem of query optimization for the virtual approach. In this approach the number of calls to an external sources is reduced and, so the processing time is more efficient.

#### 4. Schema Matching Problem:

As the web contain the heterogeneous sources, which is stored under different formats with its source schema. The need for automated data translation mechanisms is increased. However the schema matching is used to simplify the automated process without the user intervention. Where in many cases the schema of the data in the source is similar to that of target system. As we know that the integration system is need to apply the mapping between mediated schema and source schema. In this section we design a mechanisms for the translation, where the two schemas is compared (mediated schema and source schema) using a predefined rules.

In fact, the predefined rule should be known before the matching process initiated. Where in each rule the following should be considered:

```

<article>
<title > Test Page .....
<authors>
<author> A. solieman <author>
<author> S. Ahmad <author>
<author> F. mohamad <author>
</authors>

<sections>
<section> <section 1>
<title> Introduction </ title>
<body>
<parag> hello this .... </parag>
</body> </section 1> </section>

<section> <section 1>
<title> for translate ... </title>
<body>
<parag> In this section ... </ parag>
<parag> In order ... </ parag>
</body>
</section 1> </section>

<section> <section 2>
<picture> some pic </Picture>
<caption> a DTD for ... </caption>
</section 2> </section> </sections>
</article>

```

- 1) Define possible component matching between the two schemas.
- 2) Provide tools for translation of the data between them.

The rule is used to find for each component in the mediated schema a component in the source schema. And if it succeeds, the data translation is done automatically. Otherwise the user can add a new rules to describe the matching process.

An example of schema matching with a predefined rule is as follows:

Consider an XML DTD (Data Type Definition) in Figure 5, XML document in Figure 6, And the OO Database schema in Figure 7. And we would like to translate the related documents of the DTD in Figure 5 to an instance of OO Database schema as shown in Figure 8 .

```

class Article public type
    tuple ( title      : string,
           author     : list (author),
           sections   : set (section))
class Author : sting ;
class section public type
    tuple ( section 1 : tuple ( title  : string
                               ,
                               body   : set
                               (string)),
           section 2 : tuple ( pic   : string ,
                               cap   : string)

```

**Figure 7: OO database Schema**

For *section* elements: the DTD describe the union type in the OO schema and the most

similar is section class. In this class three attribute is there, where in the first two elements the name and structure is same, and in the third is *tag*. In the *tag* it indicate in which two alternatives is used. Finally we can conclude that the matching process can be applied.

- For the *author* elements in the DTD and *author* list in the OO Database, the translation process is done by translate the *author* elements

- individually and grouping then into a list.

In schema matching a set of rules is necessary to perform the translation. As mentioned above the matching rules is used to produce a target OO Database instance and this rule can be predefined (built-in) or a user can input a new rule to be added incrementally to the predefined rules.

#### 4. Conclusion and outstanding problem

Since data integration systems are designed for online querying of heterogeneous information sources, they have two other important characteristics. First, it is important to optimize the time to the initial answers to the query, rather than to minimize the total work of the system.

The hybrid approach proposed in this report is trying to overcome some drawbacks of the other approaches used in other information integration systems. Where the optimization of the running time for

query processing is reduced as an effect of reducing the number of request to the external web sources.

As we know that the integration system is need to apply the mapping between mediated schema and source schema. However, a design of mechanisms for the translation from one schema to another is demonstrated. Which is depending on comparing a mediated schema with a source schema using predefined rules to produce a target instance.

Finally, Many outstanding problem that need to be investigate, where in the data integration system the need for creation source description is essential, there is a problem on the design a framework on creation of the source description process. Also in the Information integration process the handling of violation of specific schema is not discussed and more work in this area should be done. Another problem such as in the context of extracting data or tuples from web sources, the current research is emphasizing on the syntactical structure of the web pages rather than on semantics, where the semantics extraction from the web sources is need more investigation.

#### References

- [1] Alon Y. Levy, Anand Rajaraman and Joann J. Ordille. *Querying Heterogeneous Information Sources Using Source Descriptions*. Proceedings of the 22nd International Conference on Very Large Databases, VLDB-96, Bombay, India, September, 1996
- [2] D. Konopnicki and O. Shmueli. *W3QS: A query system for the World Wide Web*. In Proceedings of the Twenty First International Conference on Very Large Data Bases, pp. 54-65, 1995.
- [3] L. Liu, C. Pu, and Y. Lee. *An adaptive approach to query mediation across heterogeneous information sources*. In Proceedings of International Conference on Cooperative Information Systems (CoopIS), pp. 144-156, Brussels, Belgium, June 1996.
- [4] D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. *Querying semistructured heterogeneous information*. Technical report, Stanford University, December 1995.

- [5] S. Chawathe, H. Garcia-Molina, J. Hammer, K. Ireland, Y. Papakonstantinou, J. Ullman, and J. Widom. *The TSIMMIS Project: Integration of Heterogeneous Information Sources*. In Proceedings of IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.
- [6] Naveen Ashish and Craig A. Knoblock. *Semi-automatic wrapper generation for internet information sources*. In Proceedings of the Second IFCIS International Conference on Cooperative Information Systems (CoopIS), Charleston, SC, 1997.
- [7] Ashish, N., and Knoblock, C. *Wrapper Generation for Semi-structured Internet Sources*. SIGMOD Record 26(4):8-15, 1997
- [8] Craig A. Knoblock, Steven Minton, Jose Luis Ambite, Pragnesh Jay Modi Naveen Ashish, Ion Muslea, Andrew G. Philpot, and Sheila Tejada. *Modeling web sources for information integration*. In Proc. Fifteenth National Conference on Artificial Intelligence, 1998.
- [9] Jose-Luis Ambite, Naveen Ashish, Craig A. Knoblock, Steven Minton, Pragnesh J. Modi, Ion Muslea, Andrew Philpot, and Sheila Tejada *ARIADNE: A System for Constructing Mediators for Internet Sources* ACM SIGMOD International Conference on Management of Data, 1998.
- [10] Genesereth, M. R., Keller, A. M. & Duschka, O. *Infomaster: An Information Integration System*, in *Proc. of 1997 ACM SIGMOD Conference*, May 1997.
- [11] Z. G. Ives, D. Florescu, M. A. Friedman, A. Y. Levy, and D. S. Weld. *An adaptive query execution system for data integration*. In Proc. of ACM SIGMOD Int. Conf. on Management of Data (SIGMOD), pp. 299-310. ACM Press, 1999.
- [12] W. Cohen. *Some practical observations on integration of Web information*. In WebDB'99 1 Workshop in conj. with ACM SIGMOD, 1999.