

# Integration of Web Data Sources: A Survey of Existing Problems

Le Quang Hieu

Institute of Computer Science  
Heinrich-Heine-University Düsseldorf  
D-40225 Düsseldorf, Germany  
lqhieu@cs.uni-duesseldorf.de

## Abstract

Integration of Web Data Sources is difficult because of the heterogeneous nature of the Web. A problem is that web data sources come in and come out frequently. Another problem is that an exact comparison between data elements in several data sources is not feasible since the data sources are owned by different organizations and therefore there are usually subtle differences in describing the same data. Or often queries need to access data sources in more than one domain. Furthermore new web standards and technologies such as XML, web services, the OWL Web Ontology Language give new opportunities and pose new challenges. In this paper, we investigate/survey on these problems and related studies as well as give our general ideas on how we could work on these problems.

## 1 Introduction

There are many web data sources available in the Internet. They are websites like Citeseer or Ebay that allow users to query information over certain criteria and answer users with data embedded in structured HTML pages. The web data sources have the following characteristics:

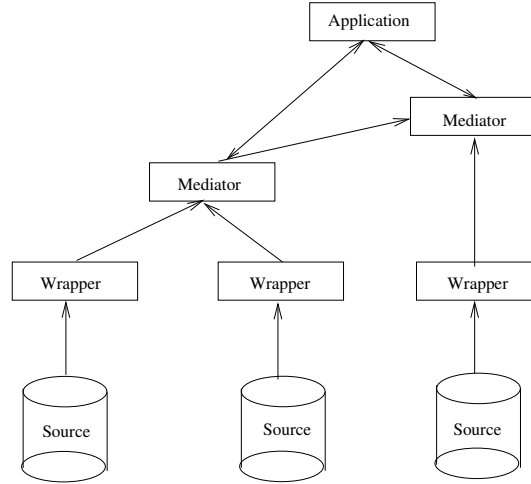
- Autonomous: Users have no control over the sources.
- Heterogeneous and Overlapping: Each source stores data in its own way, but the same kind of data is often stored in several sources; each source has different query capabilities.
- Frequently changing: Sources appear/disappear at a rapid rate; Data and layout of each source is updated continuously.
- The number of sources is very large, and it is increasing rapidly with the fast growth of the Internet.
- Distributed: sources are located all over the Internet which has an unpredictable condition. For example, connection between two points may be lost temporally or the connection speed varies from time to time. Moreover, there is no or little communication between the sources.

The integration of web data sources is to provide uniform access to multiple web data sources [9]. The integration of web data sources has many applications. It is also an active research field with many existing problems originated from the above web data sources' characteristics. The integration of web data sources is closely related to and sometimes has the similar meaning as "the integration of heterogeneous autonomous sources". For brevity, we will use "web data integration" for "integration of web data sources", and "source" for "web data source" unless explicitly indicating otherwise.

In general web data integration obviously shares many problems and techniques of traditional legacy data integration. But it has its own set of problems and techniques. The common problems in web data integration are as follows:

- Selection of architecture and data model.
- Answering queries using views and query optimization.
- Source mapping and wrapper construction.

Figure 1: A mediator architecture for web data integration



- Object similarity identification.

These above problems in turn (except of the following one) form respectively the next sections of the paper before the conclusions. In each section, we will briefly review the problem, and then give our comments on that problem. The comments are our personal views of the strength and weakness of existing solutions, or of open problems under the impacts of XML, web services, the OWL Web Ontology Language.

The two problems of answering queries using views and query optimization are not addressed in this paper. They are also the two very important problems in web data integration. An comprehensive survey for the former is [6], while a brief review for the latter can be found in [5].

## 2 Architecture and Data Model

The selection of architecture and data model play a central role for web data integration. Here we selected three systems for our following review: TSIMMIS [1], Information Manifold (IM) [10] and Ariadne [7] because they are pioneer and successful systems in this area.

There three architectures for data integration are Federated system, Mediator system [15], and Data Warehouse. It is interesting that TSIMMIS and Ariadne use the Mediator system. Though IM system has a global mediated schema, its architecture can also be considered as a mediator system since the global schema is virtual, storing no data and the IM uses wrappers to get data from sources. Figure 1 illustrates a mediator system for web data integration.

The Mediator architecture is the most suitable for web data integration since it is the most flexible architecture. It can deal better with the autonomous and frequently changing web data sources. On the other hand, both the Federated system and the Data Warehouse can not cope with the problem of adding/removing data sources frequently, and they require certain control of data sources in case of the Data Warehouse or communication between data sources in case of the Federated system.

In contrast to the use of the architecture, the three systems devise and use three different data models. TSIMMIS has the OEM (Object Exchange Model) [12], that is a simple and self-describing data model, and supports object nesting and identity. TSIMMIS has also the OEM-QL which is a SQL-like language for OEM. The data model for IM is the relational model enhanced with object-oriented features. IM use this data model along with a record for each source to describe the content of the source and the source capabilities [10]. Finally Ariadne based on SIMS uses Loom to model domains. Loom is a knowledge representation language.

The differences of data models in the three systems shows that the choice of an adequate data model is an open problem. However, one interesting note is that TSIMMIS and IM have parts of their roots in

Figure 2: Data exaction process

<p><b>The CPU List:</b></p> <p>Total: 3</p> <table> <tr> <th>Model</th><th>Speed</th><th>Price(EUR)</th></tr> <tr> <td>PIII</td><td>800MHz</td><td>60</td></tr> <tr> <td>PIII</td><td>900MHz</td><td>65</td></tr> <tr> <td>PIV</td><td>2 GHz</td><td>90</td></tr> </table> <p>Page: 1/1</p>	Model	Speed	Price(EUR)	PIII	800MHz	60	PIII	900MHz	65	PIV	2 GHz	90	<pre> &lt;html&gt; &lt;head&gt;&lt;title&gt;&lt;The CPU List&lt;/title&gt;/ead&gt; &lt;body&gt; &lt;h1&gt;The CPU List:&lt;/h1&gt; &lt;p&gt;Total: 3&lt;/p&gt; &lt;table cellpadding="8"&gt; &lt;tr&gt;&lt;td&gt;&lt;b&gt;Model&lt;/b&gt;&lt;/td&gt;&lt;td&gt;&lt;b&gt;Speed&lt;/b&gt;&lt;/td&gt; &lt;td&gt;&lt;b&gt;Price&lt;/b&gt;&lt;/td&gt;&lt;td&gt;&lt;b&gt;(EUR)&lt;/b&gt;&lt;/td&gt;&lt;/tr&gt; &lt;tr&gt;&lt;td&gt;PIII&lt;/td&gt;&lt;td&gt;800 MHz&lt;/td&gt;&lt;td&gt;60&lt;/td&gt;&lt;/tr&gt; &lt;tr&gt;&lt;td&gt;PIII&lt;/td&gt;&lt;td&gt;900 MHz&lt;/td&gt;&lt;td&gt;65&lt;/td&gt;&lt;/tr&gt; &lt;tr&gt;&lt;td&gt;PVI&lt;/td&gt;&lt;td&gt;2 GHz&lt;/td&gt;&lt;td&gt;90&lt;/td&gt;&lt;/tr&gt; &lt;/table&gt; &lt;br&gt; Page: 1/1 &lt;/body&gt; &lt;/html&gt; </pre>	<p>("PIII", 800, 60)</p> <p>("PIII", 900, 65)</p> <p>("PIV", 2000, 90)</p>
Model	Speed	Price(EUR)												
PIII	800MHz	60												
PIII	900MHz	65												
PIV	2 GHz	90												

database logic [14]. This note is also quite true for Ariadne because of its use of Loom. This fact may explain that database logic makes it easier to represent recursive relations and to re-formulate queries.

Among the three systems, IM emphasized most in the source content and capability description. However, the method used by IM may not suitable for complex sources such as Ebay. Ebay has many product categories with different sets of attributes. Therefore using the IM's method to describe Ebay contents will result in many complex views. Furthermore, the IM's method for capability descriptions is good for the sources accepting user queries in form of HTML pages, but is not applicable to sources exposed their content in form of web services like the Ebay's web service.

With the adoption of XML and especially OWL (Web Ontology Language), it is interesting to see how data models that have been created for web data integration evolve. Since XML and OWL are designed for data exchange over the Web, they have in many ways important impacts to web data integration. However OWL seems not suitable for web data integration since the data model for web data integration should be not only expressive but also simple [12] and efficient [9]. In the context of OWL adoption for the Web, a certain data model which could be evolved from the TSIMMIS's OEM may be a good one, partly because OEM is already a simple object-oriented model.

### 3 Source Mapping and Wrapper Construction

Source mapping or matching is a typical problem in data integration. That is given two schemas of two sources, find a mapping between elements of the two schemas. However, in web data integration, there is little source meta-data. Therefore sources must be learned more. The large number of web data sources makes the problem harder since manual mapping seems impractical. The source mapping should be as automated as possible. Many researches are towards that goal and they use approaches such as linguistics, constrain-based at different schema or instance level. An excellence survey of automatic schema mapping is [13].

Different from the source mapping problem, wrapper construction is a quite specific problem in web data integration. Figure 2 illustrates the problem.

Figure 2 shows that, a user asks for a list of CPUs cheaper than EUR 100. The user then gets the result as a list of CPUs in form a HTML page displayed in the left column. The HTML page is designed for humans, not for computers. The HTML page actually consists of the HTML tags, text and data shown in the middle column. To automate the extraction of data for a integration system, we need to build a wrapper that automatically parses the HTML page in the middle column to get the embedded data that are three tuples shown in the right column. As in the source mapping problem, the main problem in building wrappers is to make it as automated as possible. The reason is that there is a large

number of sources, and each source does not only frequently update its data but also its layout. There are also many approaches in the field, which use HTML-syntax analysis, Natural Language Processing, Wrapper induction, Modeling-based and so on. A good brief survey of wrapper construction is [8].

Among the approaches for source mapping, the approach in the LSD system [4] is an interesting one. In LSD, there are a meta-learner which is capable of combining the result of other learner. In that way the system may be able to extend to take advantage of other approaches.

An observation is that the increasing use of XML and OWL will ease the two problems. XML will make the process of extracting data a lot easier since XML allows the separation of data and presentation. XML and especially OWL will also reduce the difficulty in the process of schema mapping because of their self-description feature and the possibility of sharing the common data model and terminologies for a specific application domain by adopting OWL. However, the need for automatically wrapper building and source mapping remains because of the number and diversity of sources

Another observation is current methods of wrapper construction are not designed for web service data sources. On the other hand, the use of web services is increased quickly. Web sites, which make use of web services, are often large sites and store a very large amount of data such as Ebay(.com), Yahoo(.com). Among current wrapper construction approaches, the approach used in the XWRAP system [11] can be modified for web service source since it provide a framework with the separation of tasks for each source and tasks repetitive for any sources and a two-phase code generation.

## 4 Object Similarity Identification

Object similarity identification problem is that given two objects or data items of two sources, how to decide whether the two objects are "similar". This problem is quite similar to the source mapping problem, but harder. The problem is difficult for many reasons: how to define the 'similarity' meaning; the same data item is often stored in subtly different ways in two sources; the huge amount of data and so on.

There are not many researches in the problem. The two approaches frequently mentioned are [3] and [2]. The latter using textual similarity with techniques from Information Retrieval seems more suitable for web environment. One note is that the second approach is still being developed, while there are no new results for the first approach after the mentioned paper.

## 5 Conclusions

The Integration of Web Data Sources is a large research field. While it shares many problems with the traditional data integration, it also has its own set of problems because of the characteristics of the web data sources. Web data integration research is evolving with the development of the Web as well as of the Data Integration methods.

Currently web data integration only solves the problems relating to answering queries, but not with transactions between sources. The reason may be that most nowadays sources are autonomous. However, building a full featured data integration system over the Web could also a good way since it can take the advantages of Web infrastructure and matured technologies.

## References

- [1] Chawathe, S., Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J.D., Widom, J.: The TSIMMIS project: Integration of heterogeneous information sources. In: 16th Meeting of the Information Processing Society of Japan, pp. 7–18. Tokyo, Japan (1994)
- [2] Cohen, W.W.: Integration of heterogeneous databases without common domains using queries based on textual similarity. pp. 201–212 (1998)
- [3] D. McLeod D. Fang, J.H.: The identification and resolution of semantic heterogeneity. In: In Pro. First International Workshop on Interoperability in Multidatabase Systems. Kyoto, Japan (1991)
- [4] Doan, A., Domingos, P., Levy, A.Y.: Learning source description for data integration. In: WebDB (Informal Proceedings), pp. 81–86 (2000)

- [5] Florescu, D., Levy, A.Y., Mendelzon, A.O.: Database techniques for the world-wide web: A survey. *SIGMOD Record* **27**(3), 59–74 (1998)
- [6] Halevy, A.Y.: Answering queries using views: A survey. *VLDB Journal: Very Large Data Bases* **10**(4), 270–294 (2001)
- [7] Knoblock, C.A., Minton, S., Ambite, J.L., Ashish, P.J.M.N., Muslea, I., Philpot, A.G., Tejada, S.: Modeling web sources for information integration. In: *Proc. Fifteenth National Conference on Artificial Intelligence* (1998)
- [8] Laender, A.H.F., Ribeiro-Neto, B.A., da Silva, A.S., Teixeira, J.S.: A brief survey of web data extraction tools. *SIGMOD Rec.* **31**(2), 84–93 (2002). DOI <http://doi.acm.org/10.1145/565117.565137>
- [9] Levy, A.Y.: Logic-based techniques in data integration pp. 575–595 (2000)
- [10] Levy, A.Y., Rajaraman, A., Ordille, J.J.: Querying heterogeneous information sources using source descriptions. In: *Proceedings of the Twenty-second International Conference on Very Large Databases*, pp. 251–262. VLDB Endowment, Saratoga, Calif., Bombay, India (1996)
- [11] Liu, L., Pu, C., Han, W.: XWRAP: An XML-enabled wrapper construction system for web information sources. In: *ICDE*, pp. 611–621 (2000)
- [12] Papakonstantinou, Y., Garcia-Molina, H., Widom, J.: Object exchange across heterogeneous information sources. In: P.S. Yu, A.L.P. Chen (eds.) *11th Conference on Data Engineering*, pp. 251–260. IEEE Computer Society, Taipei, Taiwan (1995)
- [13] Rahm, E., Bernstein, P.A.: A survey of approaches to automatic schema matching. *VLDB Journal: Very Large Data Bases* **10**(4), 334–350 (2001)
- [14] Ullman, J.D.: Information integration using logical views. In: *ICDT '97: Proceedings of the 6th International Conference on Database Theory*, pp. 19–40. Springer-Verlag, London, UK (1997)
- [15] Wiederhold, G.: Mediators in the architecture of future information systems. *IEEE Computer* **25**(3), 38–49 (1992)