# Integrating heterogeneous data sources with temporal constraints using wrappers

Francisco Araque

E.T.S.I. Informática
University of Granada
18071 – Granada, Spain
faraque@ugr.es

**Abstract.** This paper focuses on integrating existing information sources, available via Web. We present our work related to integrating data extracted from semi-structured information sources maintaining temporal consistency of the extracted data, and monitoring the web in accordance with the user temporal requirements. The metadata about information content that is implicit in the original web pages will be extracted and encoded explicitly as XML tags in the wrapped documents. We use different approaches to maintain temporal coherency of data gathered from web sources.

## 1 Introduction

The extraordinary growth of the Internet and World Wide Web has been fueled by the ability it gives content providers to easily and cheaply publishes and distribute electronic documents. Most of the web information sources are created and maintained autonomously and each others services independently.

On the one hand, the integration of heterogeneous information sources is an important data engineering research issue [6], [8]. Data integration is a central issue in many areas to facilitate the access and manipulation of the highly distributed, heterogeneous, and dynamic collection of information sources.

An extended scheme to process information integration requirements using wrappers and temporal properties of information sources is presented. Many information sources, have their own information delivery schedules. In such cases, data arrival time is predetermined or predictable. If we use the data arrival properties of such underlying information sources, the system can derive more appropriate rules and check the consistency of user requirements more accurately. The problem facing the user now is not that the information he seeks is not available, but that it is not easy for him to extract exactly what he needs from what is available. If it is not available information about how data sources change throughout time, we use different approaches to maintain temporal coherency of data gathered from web sources.

The metadata about information content that are implicit in the original web pages will be extracted and encoded explicitly as XML tags in the wrapped documents. The final task is to generate and construct a wrapper appropriate to each Web Source from

the specification provided by the wrapper developer. This specification will include the temporal restrictions specified by the user (frequency of extraction, temporal constraints, etc).

This concepts have been implemented in a tool called *DETC* (Data Extraction with Temporal Constraints), a system for extracting and integrating data from semi-structured web sources. DETC enables users to rapidly create wrappers with temporal constraints for the Web. Using DETC's modeling tools, an application developer starts with a set of web sources -semi-structured HTML pages, which may be located at multiple web sites - and creates a unified view of these sources.

We think that in information extraction process it is necessary to maintain the temporal consistency [1], [2], [3], [10] of the data that the user needs in the sense put forward in real-time applications. In real-time applications [9], the state of the environment as perceived by the controlling system must be consistent with the actual state of the environment being controlled. In this case we employ algorithms used to maintain coherency between a data source and cached copies of the data [10]. These algorithms try to minimize the number of times that the client has to connect to the server. Besides, we use this technique to maintain the temporal consistency between the data extracted from web information sources and the data loaded in the Data Warehouse (DW) according to data warehouse designer temporal requirements [2].

## 2 Information retrieval with temporal constraints

We define a wrapper interface to specify the capability of Web Sources and extend a wrapper generation toolkit [4], [5] with graphical interfaces to specify the capability of sources and the functionality of the wrapper. The toolkit provides a graphical interface to specify the capabilities of the sources and to define a simple mapping translation from data at the source level to data at the user level. Also, we can define when we want to extract data from the source according to temporal constraints specified by the user. We have extended W4F [5] with new functionalities. First, we have developed a WysiWig interface in which the HTML document is annotated in a way that the user can get the HTML path by selecting the appropriate text (select the text he/she wants to extract, as many times as necessary, and then copy it to the interface tool with ctrl+c). Second, we have added the algorithms presented in [10].

We use Time-To-Live (TTL) values, attached to cached objects (HTML pages) [10]. Upon its expiration, the source of the object can be contacted to update the page. For minimizing the incurred network overheads, the value of TTL must be high. But a low TTL value may compromise temporal consistency. Thus any TTL value must be judged depending on two factors - how well the cache consistency is maintained, and how often the remote servers are polled. Ideally, we must dynamically update this TTL value using an algorithm that decides the value depending on the present and past rates of source changes, with the goal of keeping remote requests to a minimum while maintaining the needed temporal accuracy of the data.

## 3 Motivation example

This section shows a simple integration scenario. It defines a new information delivery channel formed by integrating three web-based information sources. Here, we assume the existence of the following three information sources: weather.lycos.com, www.weather.com and weather.yahoo.com. This requirement can be met by defining a new information delivery channel, which integrates the three information sources, polls the sources searching alteration in data and sends messages periodically according to user requirements.

The three sources have the same functioning. They offer weather related information. This Web Source supports a number of query bindings, e.g. StateName, CityName. We want to extract some piece of information from every source and create a unified view of these sources (figure 1). Each source can have different temporal characteristics attached to its data.

If we do not know when the data are going to change in the source data, we can use the DETC tool with an adaptative TTL to poll the data source (web page with the data) in order obtain the most recent data and decide what to do with it (to discard it, to load it in the DW, to store in the DB, etc).

```
.....
<weatherAtWeather>
<Time>
      9:30 PM
</Time>
<Date>
      Sunday, December 8, 2002
</Date>
<City>
      Granada, Spain
      <Tomorrow>
               Mon
               <Low>34 </Low>
               <High>42</High>
      </Tomorrow>
      <Temp>
               39
      </Temp>
</City>
</weatherAtWeather>
```

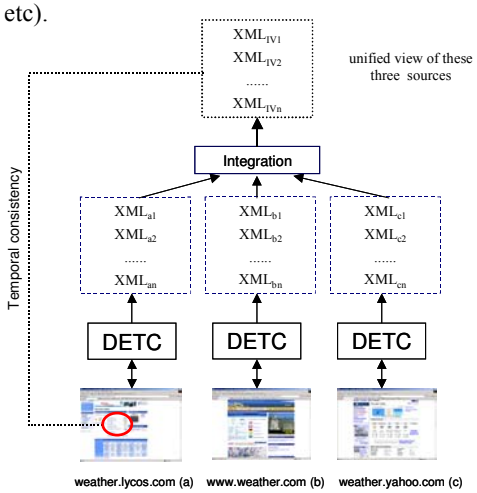


Figure 3. Integration using DETC

Of course, we can run as many wrappers as necessary depending on the number of sources that we want to poll in order to realize the integration process. Any wrapper can be parameterized with different values in accordance with the characteristics of the data source and the refreshment strategy chosen by the user. All the process (in our case all wrappers polling different web pages) is controlled by master program.

After the data has been extracted from every source, you can integrate the XML files coming from different sources in a single XML unified view. To realize this integration you must: use the same XML mappings for the three sources, or realize a new mapping from every XML template coming from every source to the new XML global template. Once you have defined the wrappers with temporal constraints and

the mappings, the integration process is completed in an automatic way. If the sources change at different speed, the most restrictive approach is adopted. That is to say, we integrate according to the source that evolves more slowly. Before integration we can transform the data if necessary, for example, we can convert the temperature from $^{o}$F to $^{o}$C range. The output in XML according to the mapping specified by the user corresponds with the XML file (unified view of the three sources) of figure 1.

## 5 Conclusion and future work

We can use the output of the wrapper (this data has been extracted with the restriction of temporal consistency specified by the user) for several purposes: to deliver to the user or client (the main finality), to store in a database for future use (for example to query, also this step requires a conversion from XML to the native format of the database), to integrate with other information to obtain a unificated vision of one or more sources, to load a data warehouse for future analysis or data mining and to notify to the user when a specific value (or threshold) has been reached. As future work, we are working in order to use ontology to carry out the integration process. Moreover, we are working to incorporate valid time and transaction time in the sense put forward in [7]. This work has been supported by the Spanish Research Program PRONTIC under project TIC2000-1723-C02-02.

## References

1. Araque, F, Samos. J., 1999: External Schemas in Real-Time Object-Oriented Databases. 20th IEEE Real-Time Systems Symposium, WIP Proceedings, pp. 105-109 (Phoenix, AZ, 12/1999).
2. Araque, F., Samos, J. Data warehouse refreshment maintaining temporal consistency. 5th International Conference on Enterprise Information Systems ICEIS'03. Angers - France - 23-26 April, 2003.
3. Araque, F., 2002: Data Warehousing with regard to temporal characteristics of the data source. IADIS WWW/Internet Conference. 13-15 November, 2002. Lisboa, Portugal.
4. Araque, F., 2002: Personalized data extraction with temporal constraints. IADIS WWW/Internet Conference. 13-15 November, 2002. Lisboa, Portugal.
5. Bhandari, D. Extraction Of Web Information Using W4F Wrapper Factory and XML-QL Query Language. Technical Report, Universitu of Pennsylvania, 1999.
6. Chawathe S., Garcia-Molina H., Hammer J., Ireland K., Papakonstantinou Y., Ullman J., and Widom J.: "The TSIMMIS Project : Integration of Heterogeneous Information Sources", IPSJ Conference, pp. 7-18, Tokyo, Japan, October 1994.
7. Fabio Grandi, Federica Mandreoli, The Valid Web: an XML/XSL Infrastructure for Temporal Management of Web Documents, Proc. ADVIS 2000 - Intl' Conf. on Advances in Information Systems , Izmir, Turkey, October 2000, LNCS 1909, © Springer-Verlag, Berlin, 2000, pp. 294-303.
8. M. Haller, B. Pröll, W. Retschitzegger, A M. Tjoa, R. R. Wagner. Integrating Heterogeneous Tourism Information in TIScover - The MIRO-Web Approach, , Information and Communication Technologies in Tourism, ENTER 2000, Barcelona, April 26-28, 2000, ISBN 3-211-83483-4.
9. Kao, B., Garcia-Molina, H., 1995: "An Overview of Real-Time Database Systems." In S. Son (Ed.), Advances in Real-Time Systems, chapter 19. Prentice Hall, 1995.
10. Srinivasan, Chao Liang, and K. Ramamritham: Maintaining Temporal Coherency of Virtual Warehouses. The 19th IEEE Real-Time Systems Symposium (RTSS98), Madrid, Spain, Dec 2-4, 1998.