ARIADNE: A System for Constructing Mediators for Internet Sources *

José Luis Ambite, Naveen Ashish, Greg Barish, Craig A. Knoblock, Steven Minton, Pragnesh J. Modi, Ion Muslea, Andrew Philpot and Sheila Tejada

Information Sciences Institute, Integrated Media Systems Center and Department of Computer Science

University of Southern California

{ambite,ashish,barish,knoblock,minton,modi,muslea,philpot,tejada}@isi.edu

http://www.isi.edu/ariadne

Abstract

The Web is based on a browsing paradigm that makes it difficult to retrieve and integrate data from multiple sites. Today, the only way to achieve this integration is by building specialized applications, which are time-consuming to develop and difficult to maintain. We are addressing this problem by creating the technology and tools for rapidly constructing information mediators that extract, query, and integrate data from web sources. The resulting system, called Ariadne, makes it feasible to rapidly build information mediators that access existing web sources.

1 Introduction

Ariadne is a system for building mediators that can gather and integrate information from multiple Internet sources. Recently, information integration from multiple sources has been done using a mediator approach. Examples include SIMS [3], Information Manifold [7], and TSIMMIS [6]. In these approaches, the mediator is aware of the information present in different sources and retrieves information from them through a 'wrapper' around each source. SIMS has been used to integrate information from heterogeneous database systems. Ariadne is a natural extension of the SIMS information integration system to the Web environment. By building wrappers around Web sources, the system can retrieve information from them in a database-like manner. In this way, Ariadne can extend the SIMS database mediator approach to the realm of Web sources. However, new problems are posed by the fact that data is being obtained from Web sources rather than databases. It is these problems that we have primarily focused on in the ARIADNE project. Our approach is to take several related Web sources in a particular domain of interest (e.g., finance, government, or real-estate) and provide the tools to rapidly construct a new application that provides integrated access to them.

For example, we can use Ariadne to provide integrated access to multiple Web sources that provide information on countries in the world (see Figure 1). An excellent Web source is the CIA World Fact Book,[1] which provides information on the geography, economy, government, etc., of every country. Another interesting source is the Yahoo listing of countries by region from where we can obtain information such as what countries are in Europe, the Pacific Rim, etc. A third relevant source is the on-line listing of Heads of state and cabinet members of all foreign governments. A user could query a mediator that provides access to the above sources to answer queries such as "Find the cabinet members of governments of countries in the Pacific Rim that have a defense expenditure exceeding $10 billion." The mediator would determine what sources can be used to answer the query, retrieve information from these sources, and present the integrated results to the user.

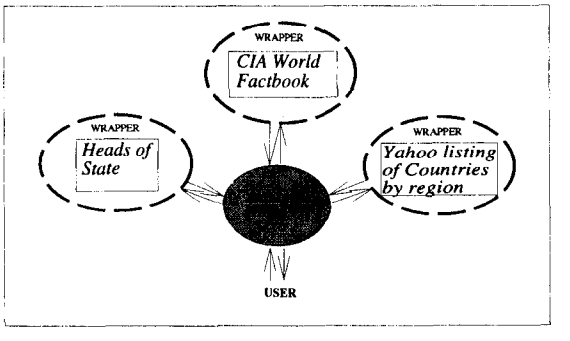

Figure 1: Integrated access to multiple Web sources through a mediator

2 System Capabilities

Ariadne has various capabilities used to rapidly construct new applications from available Web sources. The capabilities and components of the system are described below.

* This work is supported in part by the University of Southern California Integrated Media Systems Center (IMSC) - a National Science Foundation Engineering Research Center, by the Rome Laboratory of the Air Force Systems Command and the Defense Advanced Research Projects Agency (DARPA) under contract number F30602-97-2-0352, by the Defense Logistics Agency, DARPA, and Fort Huachuca under contract number DABT63-96-C-0066, and by a research grant from General Dynamics Information Systems. The views and conclusions contained in this paper are the authors' and should not be interpreted as representing the official opinion or policy of any of the above organizations or any person connected with them.

SIGMOD '98 Seattle, WA, USA

[1] http://www.odci.gov/cia/publications/nsolo/wfb-all.htm

2.1 Information Modeling

In each mediator application, we provide integrated access to a specific set of Web sources. Each source provides different but related information. The mediator is aware of exactly what information is provided by each individual source. In this way, the user is shielded from this information and is presented with an integrated view of the data. We use the LOOM knowledge representation system [8] for modeling data. A *domain* model that represents the integrated view of the data from the different sources is built in LOOM and the user poses queries against this model. We use this domain model to describe the contents of the individual sources, which allows the system to determine the best way to integrate the available sources to answer user queries.

2.2 Query Planning

The mediator needs to generate a plan that efficiently computes the answer to a user query from the relevant (Web) information sources. This plan is composed of data retrieval actions on information sources and operations on the data (such as those of the relational algebra: join, selection, etc).

Query planning for mediators in a Web-based environment presents particular challenges. We have focussed on efficient query planning techniques and addressing the limited query capabilities of web sources. First, retrieving data from web sources and transmiting data over the internet can be very time consuming, thus generating efficient plans becomes even more critical. For an efficient execution, the plan has to specify which data processing operations are going to be needed, in what order, and from which sources each different piece of information should be obtained. The highly combinatorial nature of query planning in mediators arises from these two independent sources of complexity, namely, the ordering of data processing operators and the selection of relevant information sources for terms in a given query. The Ariadne query planner [2] was designed to efficienlty generate high-quality plans while combining both source selection and traditional query optimization. Our approach, based on [1], is to generate a possibly suboptimal initial plan and then iteratively transform it by applying a set of declarative rewriting rules. Also, sometimes a plan can use information generated at run-time to determine which source to access. Our query planner uses such run-time information to minimize the number of sources accesed [5].

A second challenge arises from the fact that the capabilities of Web sources are often more limited than those of traditional database systems. In particular, many Web sources impose additional constraints on retrieval such as binding pattern constraints. Binding patterns for a relation indicate that the values of some attributes must be given in order to obtain the rest of the attributes. For example, consider a stock quote web server in which the system needs to provide the ticker symbol for a company in order to retrieve the page which has the stock value, trade volume, etc[2]. Our planner also takes into account the dependencies imposed by binding pattern constraints.

2.3 Wrappers

An essential component of the mediator architecture is wrappers around individual Web sources. The wrappers allow querying the web sources in a database-like manner (for example, using SQL). Wrappers for Web sources interpret a query from the mediator, fetch the relevant pages from the source, extract the requested information from the retrieved HTML pages and return the data to the mediator. However, it is impractical to construct wrappers for Web sources by hand because there are too many sources available and they change too fast. As a part of Ariadne, we have developed a machine learning algorithm [9] that is able to semi-automatically generate wrappers for Web sources. Whenever a new web source of interest has to be incorporated into an mediator application, we can build a wrapper for it with minimal effort using the wrapper generation toolkit [4, 9].

[2] Such a server is: http://www.quote.com/

2.4 Caching to Improve Performance

Retrieving data from Web sources is time consuming. At present the time to retrieve an individual page via HTTP is quite high (typically 2–5s). Thus answering queries that involve fetching a large number of pages over the internet can take an unacceptably long time. A considerable improvement in performance can be obtained if the mediator is able to cache frequently used data. We have incorporated a caching mechanism in the system that selectively caches frequently accessed information after analyzing the pattern of queries sent to the mediator. We first select classes of information that are useful to cache by analyzing previous user queries. We then cache the data in these classes in a database local to the mediator and define the cached classes as auxiliary information sources that the mediator can access.

2.5 Object Identification across Sources

A problem in integrating data across multiple sources is that the same objects or entities may appear differently in different sources. For example, to provide access to multiple years of the CIA World Factbook, we need to handle the fact that the name of the country "Micronesia" changes to "Micronesia_Federated_States_of" from 1995 to 1996 in the Factbook. In order to handle queries such as, what is the GNP of Micronesia from 1992 to 1996 we need to resolve these inconsistencies. Another example is when the same object is referred to by different names, such as when "International Business Machines", "IBM" or "IBM Corp." all actually refer to the same entity. Ariadne uses mapping tables and mapping functions for determining when two terms refer to the same object so that it can automatically resolve this type of inconsistency.

3 Status and Demonstration

Ariadne has been successfully used to integrate Web sources in different domains. In this system demonstration, we will show how we can query multiple sources in the two domains described below. We will also show the kind of data models built, examples of information gathering plans generated, how wrappers were generated for individual Web sources, and instances of data cached to improve performance.

3.1 Geopolitical Information Mediator

Ariadne has been used to provide integrated access to the multiple Web sources that provide information on countries in the world, which we mentioned in the Introduction. A user is able to ask queries such as queries such as "Find the

religions, ethnic groups and heads of state of all Asian countries that have a population exceeding 200 million." The mediator determines what sources can be used to answer the query, retrieves information from these sources, and presents the integrated result to the user.

3.2 Geographic Information Display Mediator

Ariadne has been used to provide integrated access to multiple Web sources that provide geographic information. Consider three heterogeneous Web sources that provide geographic information about various US cities. 1) The Tiger Map Server that provides detailed maps of each US city, 2) The Zagats Restaurants Reviews site - Web source that provides for each US city information about restaurants such as name, address, cuisine type, rating, reviews etc., and 3) The GeoCoder Web source that returns latitude and longitude of a location given the postal address.

Given integrated access to the above sources, we can obtain interesting information such as "What are the Italian restaurants in a given geographic area of Los Angeles?". The user specifies the geographic area by graphically selecting a bounding box on a map of Los Angeles on the (independently developed) user interface. The mediator places the restaurants that satisfy the user query at the appropriate coordinates in the selected area on the map.

4 Conclusion

A wealth of information can be obtained by integrating related sources of information about a particular subject already available on the Web. Ariadne demonstrates how we can rapidly build mediators to integrate Web sources with minimal effort.

References

[1] José Luis Ambite and Craig A. Knoblock. Planning by rewriting: Efficiently generating high quality plans. In *Proceedings of AAAI-97, The National Conference on Artificial Intelligence*, Providence, RI, 1997.

[2] José Luis Ambite and Craig A. Knoblock. Flexible and scalable query planning in distributed and heterogeneous environments. In *The Fourth International Conference on Artificial Intelligence Planning Systems (AIPS-98)*, Pittsburgh, June, 1998.

[3] Yigal Arens, Craig A. Knoblock, and Wei-Min Shen. Query reformulation for dynamic information integration. *Journal of Intelligent Information Systems, Special Issue on Intelligent Information Integration*, 6(2/3):99–130, 1996.

[4] Naveen Ashish and Craig A. Knoblock. Semi-automatic wrapper generation for internet information sources. In *Proccedings of the Second IFCIS International Conference on Cooperative Information Systems (CoopIS)*, Charleston, SC, 1997.

[5] Naveen Ashish, Craig A. Knoblock, and Alon Y. Levy. Information gathering plans with sensing actions. In *Proceedings of the Fourth European Conference on Planning*, Toulouse, France, 1997.

[6] Joachim Hammer, Hector Garcia-Molina, Kelly Ireland, Yannis Papakonstantinou, Jeffrey Ullman, and Jennifer Widom. Information translation, mediation, and mosaic-based browsing in the tsimmis system. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, San Jose, CA, 1995.

[7] Thomas Kirk, Alon Y. Levy, Yehoshua Sagiv, and Divesh Srivastava. The information manifold. In *Working Notes of the AAAI Spring Symposium on Information Gathering in Heterogeneous, Distributed Environments*, Technical Report SS-95-08, AAAI Press, Menlo Park, CA, 1995.

[8] Robert MacGregor. A deductive pattern matcher. In *Proceedings of AAAI-88, The National Conference on Artificial Intelligence*, St.Paul, MN, 1988.

[9] Ion Muslea, Steven Minton, and Craig A. Knoblock. Automated wrapper generation for semi-structured, web-based information sources. Submitted to AAAI-98, 1998.