AFRL-SR-BL-TR-98-

0302

REPORT DOCUMENTATION P

18

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
11/24/97

3. REPORT TYPE AND DATES COVERED
Final Technical Report:9/30/93-9/29/97

4. TITLE AND SUBTITLE
Building Information Servers
SIMS-AA Technical Report

5. FUNDING NUMBERS
C=F49620-93-1-0594

6. AUTHOR(S)
Craig A.Knoblock, William Swartout, and Sheila Tejada

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
USC INFORMATION SCIENCES INSTITUTE
4676 ADMIRALTY WAY
MARINA DEL REY, CA 90292-6695

8. PERFORMING ORGANIZATON REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)
Defense Advanced Research Projects Agency
3701 No.Fairfax Drive
Arlington, VA 22203

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

19980414 082

11. SUPPLEMENTARY NOTES

12A. DISTRIBUTION/AVAILABILITY STATEMENT
UNCLASSIFIED/UNLIMITED

12B. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

This research addressed the problem of determining the relationships among multiple, diverse information sources in order to support the integration of data from these sources. In general, to integrate data from multiple sources requires a model of the precise relationships between the sources. Constructing such a model by hand is a difficult and time-consuming process. The relationships captured in a model describe the type of overlap between data instances in different sources. In this work data mining techniques were used to determine these relationships by comparing the data instances between sources. A related problem is that data instances can exist in different formats across several sources, e.g. IBM may be abbreviated as IBM in one source and appear as International Business Machines in another source. This work addressed this problem by developing techniques for automatically determining the mapping between names used in different sources. These integration techniques were use in conjunction with the SIMS information mediator, allowing SIMS to correctly and efficiently integrate data across several sources that contained data instances appearing in multiple formats.

14. SUBJECT TERMS
information integration, mediator, data mining, source models, SIMS

15. NUMBER OF PAGES
1

16. PRICE CODE

17. SECURITY CLASSIFICTION OF REPORT
UNCLASSIFIED

18. SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED

19. SECURITY CLASSIFICATION OF ABSTRACT
UNCLASSIFIED

20. LIMITATION OF ABSTRACT
UNLIMITED

Final Report: SIMS-AA

Contract Number: F49620-93-1-0594

Dates: 9/30/96 - 9/22/97

Craig Knoblock, Project Leader
USC / Information Sciences Institute
4676 Admiralty Way, Suite 1001
Marina del Rey, CA 90292
E-Mail: knoblock@ isi.edu
ph: (310)822-1511
fax: (310)823-6714

AASERT 97 Final Report

This AASERT research was driven by the need to automate the integration of information from multiple diverse information sources available on the internet, in order to efficiently answer user queries. For example, if the user was interested in finding out the annual incomes of all the computer companies whose current stock price is greater than $100, this would involve accessing and integrating information from the Securities and Exchange Commission (SEC) web server, which has company annual reports, and a current stock quote server available on the web. To manually perform this task a user would first retrieve the annual incomes of all of the computer companies from the SEC web site, and then retrieve the current stock price of each of these companies from the stock quote server and check if the price is greater than $100. This task would require a significant amount of effort to perform on the part of the user, especially when retrieving large amounts of data. The user would also need to possess a great deal of knowledge about how to access the data contained in the sources, as well as how to integrate the two sets of data. A more desirable approach would be to provide the user with a single interface that allows access to multiple information sources, abstracting away the need for the user to know the location or query access methods of any particular source.

The SIMS information broker was designed to provide the user with such an interface. Therefore, as shown by this example, in order to properly access and integrate the data from independent heterogeneous sources the system must have knowledge about the relationships of the data contained in each of these sources, as well as the relationships between the sources. SIMS captures this type of relationship information in the form of a model, called a domain model. The types of relationships captured in the model describe the amount of overlapping data instances shared between the sources. Presently, the domain model is manually generated by human experts who are familiar with the data stored in the sources. To automatically generate domain models, datamining techniques are used to determine which data instances appear in multiple sources, e.g. which (if any) companies from the SEC web site, like the computer company IBM, also have stock quote information in the stock quote server. Once the overlapping data instances are determined, the relationship between the data in the sources can be modeled in the domain model as either subset, superset, equality, overlapping or association.

Since datamining techniques determine these relationships by comparing the data instances between sources to discover which data instances are shared between the sources; potentially, all instances from one source could be compared with all of the instances from an other source. In order to constrain the number of comparisons each source is first mined for the properties or features of its data, such as its type, length and range; then , only instances which are have compatible properties are compared. The experimental results showed that this technique reduced the number of comparisons performed when mining the source for the model relationships, and that in some cases the number of comparisons were reduced by 70 percent.

A special case for information integration is when data instances can exist in different formats across several sources, e.g. the company IBM can appear as International Business Machines in another source. In this case, constructing a model that represents only the amount of overlap between sources is not sufficient to properly integrate the retrieved data. Information relating each specific pair of corresponding data instances must also be captured, e.g. (IBM, International Business Machines). This information is stored in a mapping table which is modeled as an information source with overlapping relationships between the sources for which it contains mapping information. In other words, the mapping table source has subset relationships with the SEC and stock server sources. This integration technique has allowed SIMS to properly and efficiently integrate data across several sources that contained data instances appearing in multiples formats.

Sheila Tejada's grades are satisfactory. Her GPA is 3.6