Exploiting Secondary Sources for Automatic Object Consolidation

Martin Michalowski, Snehal Thakkar, and Craig A. Knoblock
University of Southern California
Information Sciences Institute and Department of Computer Science
4676 Admiralty Way
Marina del Rey, CA 90292
{martinm,thakkar,knoblock}@isi.edu

ABSTRACT

Information sources on the web are controlled by different organizations or people, utilize different text formats, and have varying inconsistencies. Therefore, any system that integrates information from different data sources must consolidate data from these sources. Data from many data sources on the web may not contain enough information to accurately consolidate the data even using state of the art object consolidation systems. We present an approach to accurately and automatically consolidate data from various data sources by utilizing a state of the art object consolidation system in conjunction with a mediator system. The mediator system is able to automatically determine which secondary sources need to be queried in cases where the object consolidation system is unable to confidently determine whether two records refer to the same entity. In turn, the object consolidation system is then able to utilize this additional information to improve the accuracy of the consolidation between datasets.

1. INTRODUCTION

Web-based information integration systems such as Information Manifold [7], InfoMaster [4], and Ariadne [6] can provide a uniform query interface to the users to query information from various web sources as well as databases. While the above mentioned systems can integrate information from various data sources, none of them completely address the issues relating to text formatting inconsistencies across several data sources. For example, two real estate web sites may refer to the same address using different text formatting. Therefore, object consolidation is essential to accurately integrate data from various data sources.

There has been some work done on consolidating data objects from various web-sites using textual similarities and transformations [1, 2, 3, 5, 8, 9, 10]. These approaches provide better consolidation results compared to the exact text matching techniques in different application domains. However, in some application domains it may be extremely difficult to consolidate records. For example, when matching names of people, it would be hard for the above-mentioned techniques to determine if “Robert Smith” and “Bob Smith” refer to the same individual. This problem can often be solved by utilizing information from different data sources on the web. For example, a web site that lists the common acronyms used for the first name may provide information that “Bob” and “Robert” are interchangeable as first names. Other examples include utilizing a geocoder to determine if two addresses are the same, utilizing historical area code changes to determine if two phone numbers are the same, and utilizing the location and officers information for different companies to determine if two companies are the same.

In this paper, we describe our work on exploiting secondary sources for automatic object consolidation. The goal of the research is to (a) provide a uniform query interface to data from different web sites, and (b) consolidate this data to eliminate duplicates and formatting inconsistencies across various data sources.

2. MOTIVATING EXAMPLE

To clarify the concepts in the paper, we first define the following terms: (1) object consolidation, (2) primary data sources, and (3) secondary data sources. Object consolidation is the process of combining records from two data sources into one record set containing one record per entity. A primary data source is one of the two initial data sources used for object consolidation. A secondary data source is any source, other then a primary data source, that can provide additional information about entities in the primary data sources.

We utilize the example domain model shown in Figure 1 to motivate the use of secondary data sources. Our example mediator system has access to various data sources to obtain restaurant, hotel, and theater information. The mediator system also has access to a Geocoder that provides geographic coordinates for a given address, and an area code updates source that provides information about area code changes. There may be some inconsistencies between different data sources. For example, the Zagat data source and the Dinesite data source may use different formats for addresses, or restaurant names.

Michalowski ’03 Washington, DC USA

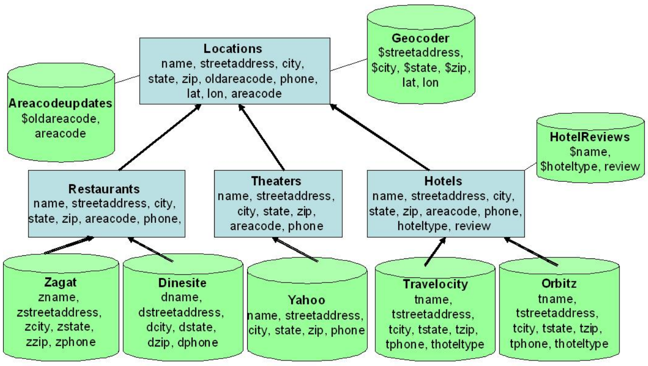
Locations
name, streetaddress, city, state, zip, oldareacode, phone, lat, lon, areacode
Geocoder
$streetaddress, $city, $state, $zip, lat, lon
Areacodeupdates
$oldareacode, areacode
HotelReviews
$name, $hoteltype, review
Restaurants
name, streetaddress, city, state, zip, areacode, phone,
Theaters
name, streetaddress, city, state, zip, areacode, phone
Hotels
name, streetaddress, city, state, zip, areacode, phone, hoteltype, review
Zagat
zname, zstreetaddress, zcity, zstate, zzip, zphone
Dinesite
dname, dstreetaddress, dcity, dstate, dzip, dphone
Yahoo
name, streetaddress, city, state, zip, phone
Travelocity
tname, tstreetaddress, tcity, tstate, tzip, tphone, thoteltype
Orbitz
tname, tstreetaddress, tcity, tstate, tzip, tphone, thoteltype

Figure 1: The Mediator's Domain Model

3. ACTIVE ATLAS: OBJECT CONSOLIDATION SYSTEM

3.1 System Overview

The Active Atlas system developed by Tejada et al. [9] is an object consolidation system consisting of two separate components: a candidate generator and a mapping learner. The system's goal is to find common entities amongst two record sets from the same domain. This process involves two stages: Firstly, transformations are applied to calculate the initial similarity scores. Secondly, mapping rules and transformation weights are learned and used in the mapping of objects between sources. The output of the system consists of two elements: (1) a set of consolidated records and (2) transformation weights and mapping rules pertaining to that specific domain.

3.2 Open Research Problem

A difficult problem encountered when performing object consolidation is the degree of certainty with which matches are proposed and rejected. An object consolidation system is only as good as the labeled data it has received and is therefore limited in accuracy with respect to its classification of matches. In our research of object consolidation, we have found that there exists a "grey" area in the classification of potential matches. An object consolidation system is able to classify obvious matches and non-matches. However, a class of potential matches is also present. This class needs to be dealt with in a different and more compelling manner.

The potential matches cannot be classified with full confidence as a match yet they possess a score high enough to be considered as potentially matched. This presents the need for a secondary source to help resolve this discrepancy. A secondary source would provide the system with additional information which it could use to help in the classification of the match.

4. EXPLOITING SECONDARY SOURCES FOR OBJECT CONSOLIDATION

In this section, we describe our approach in combining a mediator system with an object consolidation system to improve the performance of both systems. The performance of the mediator system improves because it can utilize the object consolidation system to consolidate records from various data sources. The performance of the object consolidation system also improves by utilizing information, provided by the mediator, from the secondary sources. The key challenges to this approach are as follows: (1) the mediator needs to determine when to query secondary data sources, and (2) the mediator needs to determine which secondary data sources should be queried.

Our approach differs from traditional mediators in two ways: (1) we utilize an object consolidation system to consolidate data from various data sources, and (2) our approach automatically improves the performance of the object consolidation system by utilizing information from the secondary sources. The mediator system utilizes the domain descriptions to transform the user query into a datalog program that queries various data sources and processes the data from these sources to answer the query.

4.1 Object Consolidation in the Mediator

The object consolidation system used in our approach is a modified version of the Active Atlas System[9]. Our object consolidation component consists of two main elements, the candidate generator and the evaluator. First, the candidate generator works in the same manner as mentioned in [9]. Secondly, the evaluator consists of a matching mechanism and a mapping rule evaluator. The evaluator takes as input two data sets produced by the candidate generator, a set of transformation weights and mapping rules and produces a set of proposed matches. This element does not learn the transformation weights and mapping rules but rather assumes that they have been provided elsewhere and are passed in as input. Furthermore, the mapping rule evaluator has been altered to understand mapping rules that classify proposed matches as "Maybe". This is an important addition and its relevance is discussed in Sections 4.2 and 4.3.

To present the importance of object consolidation in the mediator, we will use an example from the model presented in Figure 1 involving records from the Zagat and Dinesite sources. Both of these sources have a potentially different representation of each attribute (name, streetaddress etc..) and we must perform object consolidation to determine common entities between the two sources. If this consolidation step finds matches that potentially refer to the same entity but require more information to be certain, the mediator will query an available secondary source such as the Geocoder to obtain the required additional information. Once this information is obtained and joined to the corresponding records in the potential matches, object consolidation must be executed again.

4.2 When To Query Secondary Data Sources

As mentioned in Section 3, the Active Atlas system utilizes the user labeled examples to learn the transformation weights and mapping rules to consolidate the records from various data sources. We assume that the user has run the Active Atlas or a similar object consolidation system previously and has provided a set of transformation weights and mapping rules to consolidate data from various data sources. Our approach allows the user to specify mapping rules that can classify a match between any two records from the given data sources as "matched", "not matched", or "not sure".

When the mediator needs to consolidate data from many sources, the mediator passes the data, the corresponding transformation weights, and the corresponding mapping rules to the object consolidation component. The object consolidation component then returns the consolidated data back to the mediator. If the consolidated data contains tuples classified as "not sure" the mediator decides to obtain information from the secondary sources to improve the performance of the object consolidation system. It is better for the system to incorporate additional information only for uncertain matches as opposed to querying the secondary source for each record in both primary data sources before object consolidation for several reasons: (1) Querying secondary sources may increase latency and processing, (2) Not all records benefit from the additional information provided by the secondary sources, and (3) the additional information also increases the complexity of the object consolidation process. Next, we show how the mediator determines which secondary sources should be queried.

4.3 Which Secondary Sources Should Be Queried

The mediator utilizes domain descriptions to determine which secondary sources should be queried to help with the consolidation process. We will describe this process by going through an example consolidation process for the domain in Figure 1. While consolidating the restaurant records obtained from the Zagat and Dinesite data sources and finding that the consolidation component has classified some of the matching records as "not sure", the mediator analyzes the domain description to find that the Zagat and the Dinesite data sources provide information about the Restaurants class. The mediator further analyzes the domain description to find that there are no sources that provide more information about the Restaurants class. The mediator continues its search and finds that the Restaurant class is a subclass of the Locations class and the sources Geocoder and AreaCodeUpdates provide more information about the Locations class. Therefore, the mediator utilizes the Geocoder and the AreaCodeUpdates data sources as secondary sources.

5. FUTURE WORK AND CONCLUSIONS

We described an approach to utilize secondary sources for object consolidation. We believe that our approach can greatly enhance the accuracy of web-based object consolidation. We are evaluating different methods to pick the most promising secondary source when multiple secondary sources are available for consolidation. In some cases, data sources do not provide "nicely" segmented fields, but give descriptions composed of multiple fields concatenated together, e.g. the whole address and telephone in one string. We are working on machine learning techniques to automatically segment such data into fields by learning a grammer for the given data, e.g. divide address into street number, street name, city, state, and zip code.

We would like to incorporate a learning component into the system which computes transformation weights and mapping rules internally. We believe we can exploit the existence of additional data by incorporating this data into the examples presented to the learning component. By enhancing the examples available to the system, more independent learning would occur, potentially leading to a totally unsupervised learning component.

6. ACKNOWLEDGEMENTS

This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA) and Air Force Research Laboratory under contract/agreement numbers F30602-01-C-0197 and F30602-00-1-0504, in part by the Air Force Office of Scientific Research under grant numbers F49620-01-1-0053 and F49620-02-1-0270, in part by the United States Air Force under contract number F49620-02-C-0103, and in part by a gift from the Microsoft Corporation. The U.S. Government is authorized to reproduce and distribute reports for Governmental purposes notwithstanding any copy right annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of any of the above organizations or any person connected with them.

7. REFERENCES

[1] M. Bilenko and R. J. Mooney. Employing trainable string similarity metrics for information integration. In *Proceedings of the IJCAI Workshop on Information Integration on the Web*, Acapulco, Mexico, August 2003.

[2] W. W. Cohen. Reasoning about textual similarity in a web-based information access system. *Autonomous Agents and Multi-Agent Systems*, 2(1):65–86, 1999.

[3] W. W. Cohen and J. Richman. Learning to match and cluster large high-dimensional data sets for data integration. In *Proceedings of the Eighth ACM SIGKDD International Conference*, Edmonton, Alberta, Canada, July 2002.

[4] M. R. Genesereth, A. M. Keller, and O. M. Duschka. Infomaster: an information integration system. In *1997 ACM SIGMOD Conference*, pages 539–542, 1997.

[5] M. A. Hernndez and S. J. Stolfo. The merge/purge problem for large databases. In *Proceedings of SIGMOD-95*, pages 127–138, May 1995.

[6] C. A. Knoblock, S. Minton, J. L. Ambite, N. Ashish, I. Muslea, A. Philpot, and S. Tejada. The ariadne approach to web-based information integration. *International Journal of Cooperative Information Systems*, 10(1-2):145–169, 2001.

[7] A. Y. Levy, A. Rajaraman, and J. J. Ordille. Querying heterogeneous information sources using source descriptions. In *Proceedings of the Twenty-second International Conference on Very Large Databases*, pages 251–262, Bombay, India, 1996. VLDB Endowment, Saratoga, Calif.

[8] S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the Eighth ACM SIGKDD International Conference*, Edmonton, Alberta, Canada, July 2002.

[9] S. Tejada, C. A. Knoblock, and S. Minton. Learning domain-independent string transformation weights for high accuracy object identification. In *Proceedings of the Eighth ACM SIGKDD International Conference*, Edmonton, Alberta, Canada, July 2002.

[10] W. E. Winkler. The state of record linkage and current research problems. *Technical Report, Statistical Research Division, U.S. Bureau of the Cenus*, 1999.