Mapping Target Schemas to Source Schemas Using WordNet Hierarchies

A Thesis Proposal Presented to the
Department of Computer Science
Brigham Young University

In Partial Fulfillment of the Requirements
for the Degree of Master of Science

David Jackman

September 19, 2000

I. Introduction

As the amount of information and number of information sources continue to grow, so does the need for better data integration techniques. Integration of independent sources of information within a domain of interest would provide much greater value than any single source could ever offer. However, integrating the data in heterogeneous data sources is difficult.

Much research has been devoted to solving the problem of information integration. In the mid-1980's, research focused on integrating schemas when designing large databases ([BLN86]). In the early 1990's, this research evolved to integrating whole databases ([SL90], [TTCB+90]). Today, research initiatives are aimed at integrating information from various structured and semi-structured sources. Some attempts (e.g. [CGHP+94], [KLSS95], [Ull97], [Coh98]) integrate the data sources as queries are posed, while others (e.g. [BCV99], [BE99], [FPNB99]) integrate the data sources beforehand, offering a general schema that is queried. Regardless of when the integration takes place, the same problems arise. Arguably the greatest of these challenges is resolving semantic conflicts between two sources. The core of this problem is determining whether object or relationship sets in different sources are related and if so, then whether this relationship is equality, generalization, specialization, or some type of intersection.

Much of the difficulty in automatically resolving semantic conflicts between heterogeneous data sources arises from terminology differences in the different schemas. [BE99] uses keywords and sample values to aid in matching object sets. [Coh98] uses techniques from information retrieval research to approximate a mapping between query

terms and schema terms. But most approaches require humans to do much of the work to provide a mapping between heterogeneous schemas.

WordNet is a lexical database of the English language that has done much to further research in natural language processing. The database provides a list of all the noun, verb, and adjective word senses for each word, along with that word sense's placement in a hierarchy of words within that domain. (For example, apple is a kind of fruit, which is a kind of food.) WordNet is described in detail in [Mil95] and is available for download and review from the WordNet web site ([WN]).

This research proposes to use the WordNet lexical database to aid in automatically resolving semantic conflicts by providing possible relationships between object and relationship sets of different data sources based on the terminology of those sources and their position in the WordNet word-sense hierarchy. The integration approach in [BCV99] briefly mentions WordNet as a tool that could be used to propose possible relationships between objects sets based on field names, but does not give any details about how this would be done or its effectiveness on real data. [CA99] uses WordNet in conjunction with other thesauri to identify terminological relationships, but no details are given about its effectiveness or limitations.

II. Thesis Statement

This research will investigate the use of the WordNet lexical database to aid in making semantic associations between heterogeneous data sources. Given two schemas, WordNet information will be used to determine which object sets are most likely to be

related. Contextual information based on potential WordNet associations in target and source schemas will also be used to increase the confidence in these relationships.

III. Methods

This research is being conducted within the integration framework presented in [BE99]. The framework in [BE99] supposes that a target schema exists and that one or more source schemas are to be integrated into the target schema. The target schema and the source schemas are all assumed to be OSM model instances [EKW92]. To do the integration an injective target-to-source mapping between the object and relationship sets of the target and source OSM model instances is required.

The objective of this research is to investigate the use of WordNet to produce a mapping from a target schema to a source schema. This application framework will take as input the target and source schemas, which will be given as XML documents. The output is another XML document, which gives the possible mappings and the most likely mapping between the target and source schemas. The application framework will be entirely written in Java, and will contain the necessary code for processing the input and output files.

The particular approach taken in this research will determine possible target to source mappings through the use of two procedures, one that uses WordNet data only, and another that considers the context of each object set based on the results of the WordNet data associations. Each procedure will assign a confidence score to each mapping based on its own results, and these scores will then be used to give a final score to each mapping. The application framework will then use the final scores to determine the

highest scoring complete set of mappings, assuring each object set is only used in a single mapping.

WordNet Procedure

The WordNet procedure will consider each possible target to source mapping of object sets and calculate a WordNet score for each mapping based on the distances between those terms and the closest common parent term in the WordNet hierarchy. A smaller distance for the pair of words will attain a higher WordNet score for the pair. Many of the possible target and source term pairs will not have a common WordNet parent, so this procedure will also eliminate many of the implausible mappings (with respect to WordNet).

For the WordNet procedure to be successful, it is required that all abbreviations and acronyms in the object set labels be expanded in advance. This procedure can be automated through the use of data dictionaries, but for this research these expansions are assumed to have already been done. For object-set labels with multiple words, this procedure will use only the last noun term in the label, excluding prepositional phrases. For example, in the phrase "narrowest width in centimeters," the WordNet procedure would use "width." Some object sets may also have synonyms associated with them. When these are available, the WordNet procedure will consider these terms as well when scoring the possible mappings.

Context Procedure

The context procedure uses only those target-to-source mappings whose scores from the WordNet procedure exceed a specified threshold. This procedure calculates a score for each mapping based on the context of the mappings around it in the target and source

OSM model instances. Most object sets in the target schema will have several possible mappings based on the WordNet procedure. The most likely mapping for each object set will be the one that has a high WordNet score and a high contextual score.

For each mapping M_0, which maps target object set T_0 to source object set S_0, the context procedure considers the mappings $M_1, \ldots, M_n$, which map target object sets $T_1, \ldots, T_p$ to source object sets $S_1, \ldots, S_q$, where each of $T_1, \ldots, T_p$ is adjacent to T_0 (meaning each T_i ($1 \le i \le p$) is connected to T_0 by a single relationship set in the OSM model instance). The context score for M_0 increases for each adjacent mapping $M_1, \ldots, M_n$ that exists (given the qualifications described below). The distance between the S_0 and S_i ($1 \le i \le q$), given by the number of object sets in the shortest path between the two object sets in the OSM model instance, determines the amount of the increase. A smaller distance yields a greater increase.

Where multiple mappings exist for a single adjacent target object set, the context procedure will automatically choose only the best one. This heuristic choice will depend on several characteristics. (1) Mappings with the smallest distance in the source schema will be favored. (2) Each object set in the source schema should only be used once when calculating a context score. (3) Mappings should be less favorable considered if the cardinality of the target object set T_0 to an adjacent target object set T_i is not compatible with the cardinality of the source object set S_0 to the corresponding source object set S_j for the mapping (T_i, S_j).

Experiments

We will run several experiments to determine the effectiveness of the combined WordNet/context heuristic procedures on real data. Four integration tasks will be run.

The four applications for these tasks will be (1) music CD's, (2) genealogy, (3) travel, and (4) real estate. For each task both the combined WordNet/context heuristic procedures and a human expert will produce target-to-source mappings. Assuming the human expert is correct, the accuracy of the heuristic procedure will be judged. False drops and false positives will be analyzed and discussed. Before conducting these experiments, several "training" sets of schemas will be used to tune the scoring heuristics. We will also explore the complexity of the algorithms used in this process.

In the experiment three source database schemas will be used for each of the four applications. The source schemas will be taken from existing web pages, mimicking the approach of example-based web extraction tools (e.g. [GLSR00]). The field labels are to be taken directly from the web pages whenever they are available. For those fields that have implied labels in the web pages, suitable field names will be created, based as much as possible on the context of that field, including the terms used on any search forms or result tables that link to the pages.

A target schema for each of the four applications will be constructed from industry standard XML DTDs, many of which are given in [XML], for those domains where DTDs can be found that closely match the kinds of data that are represented in the source web pages. In this case, the field labels will be taken directly from the XML element names. Where a suitable XML DTD cannot be found for a domain, a target schema will be created by an independent party told to describe the data he or she would like to see in an integrated web site for that domain. In this case, the field labels will be taken directly from the independent descriptions.

IV. Contribution to Computer Science

This research will further the work currently being done in heterogeneous data integration by providing a means whereby more of the data integration process can be performed automatically. While the data integration process may never be completely automatic, this research will make a data-integration tool better able to suggest plausible mappings to the user by taking advantage of the natural language characteristics encoded in WordNet.

V. Delimitations of the Thesis

This research will not attempt to do the following:

- Automatic preprocessing of the schema to translate abbreviations and acronyms into terms that exist in the WordNet hierarchy.
- Explore the use of specialized word hierarchies that may be constructed in different domains.
- Explore how the techniques could be extended to languages other than English.

VI. Thesis Outline

1. Introduction (3 pages)
2. Related Work (2 pages)
3. Integration Tool Framework (15 pages)
   3.1. Overview and Program Flow
   3.2. Input Documents
   3.3. Output Documents
4. Integration Mapping Generation (25 pages)
   4.1. WordNet Procedure
   4.2. Contextual Procedure
5. Experimental Analysis and Results (10 pages)
6. Conclusions, Limitations, and Future Work (4 pages)

VII. Thesis Schedule

A tentative schedule of this thesis is as follows:

Literature Search and Reading	January – July 2000
Chapter 3	September – October 2000
Chapter 4	October – November 2000
Chapters 1 and 2	December 2000 – January 2001
Chapters 5 and 6	January – February 2001
Thesis Revision and Defense	March 2001

VIII. Bibliography

[BCV99] S. Bergamaschi, S. Castano, and M. Vincini. "Semantic Integration of Semistructured and Structured Data Sources". *SIGMOD Record* 28 (1), March 1999, pp. 54-59.

This paper describes the MOMIS approach to integration and query of multiple, heterogeneous information sources. A common thesaurus is constructed and serves as a shared

ontology for the information sources. WordNet is briefly mentioned as a way to build this common thesaurus, but no detailed information is given about how this is done or any limitations of this approach.

[BLN86] C. Batini, M. Lenzerini, and S. B. Navathe. "A Comparative Analysis of Methodologies for Database Schema Integration". *ACM Computing Surveys*. 18(4):323-364, December 1986.

This article provides a representation of the work being done in the mid-1980's in database schema integration.

[BE99] J. Biskup, and D. Embley. "Extracting Information from Heterogeneous Information Sources Using Ontologically Specified Target Views".

This paper describes the approach to database integration that is followed by this proposal. Their approach is unique in its use of custom ontologies to describe individual database schemas, and creates a target "view" of the information sources which is then mapped to the actual source objects and relationships.

[CA99] S. Castano, and V. De Antonellis. "ARTEMIS: Analysis and Reconciliation Tool Environment for Multiple Information Sources". *Proceedings of the Convegno Nazionale Sistemi di Basi di Dati Evolute* (SEBD 99): 341-356, Como, Italy, June 1999.

This paper describes the ARTEMIS project, which is "a tool environment developed to support the analyst in the process of analyzing the reconciling sets of heterogeneous data schemas". WordNet is used in conjunction with other thesauruses to identify terminological relationships, but no details are given about its effectiveness or limitations.

[CGHP+94] S. Chawathe, H. Garcia-Molina, J. Hammer, Y. Papakonstantinou, J. Ullman, and J. Widom. "The Tsimmis Project: Integration of Heterogeneous Information Sources". In *Proceedings of 100th Anniversary Meeting of the Information Processing Society of Japan*, pages 7-18, Tokyo, Japan, October 1994

This paper gives an overview of the TSIMMIS project, one of the pioneering research projects in information integration. Their approach involves the creation of custom translators and mediators for data sources, which classify data, translate queries, and filter result sets.

[Coh98] W.W. Cohen. "The WHIRL Approach to Integration: An Overview". In *Proceedings of the AAAI Workshop on AI and Information Integration*, Madison, Wisconsin, July 1998.

This paper describes an information integration project that uses ranked retrieval methodology from information retrieval research to automatically approximate the translation of a query to use the terms and keys for each information source.

[EKW92] D.W. Embley, B.D. Kurtz, and S.N. Woodfield. *Object-oriented Systems Analysis: A Model-Driven Approach.* Prentice Hall, Englewood Cliffs, New Jersey, 1992.

This provides a reference for OSM modeling, which is used by this research.

[FPNB99] J. Fowler, B. Perry, M. Nodine, and B. Bargmeyer. "Agent-Based Semantic Interoperability in InfoSleuth". *SIGMOD Record*, 28(1):60-67, March 1999.

InfoSleuth is a distributed agent architecture that provides for semantic interchange through the creation of a single ontology that expresses the concepts and relationships of the application domain in high-level terms that are then translated to the low-level types for each database schema.

[GLSR00] Paulo B. Golgher, Alberto H.F. Laender, Altigran S. da Silva, Berthier Ribeiro-Neto. "An Example-Based Environment for Wrapper Generation". *Proceedings of the International Conference on the World Wide Web and Conceptual Modeling*, Salt Lake City, Utah, 9-12 October 2000 (to appear).

This paper describes a tool that builds web extraction wrappers by example, based on positioning and formatting of text in the target web pages.

[Mil95] G. Miller. "WordNet: a lexical database for English." In: *Communications of the ACM* 38 (11), November 1995, pp. 39-41.

This is the original paper the introduced the WordNet research project.

[KLSS95] Thomas Kirk, Alon Y. Levy, Yehoshua Sagiv, and Divesh Srivastava. "The Information Manifold". In *Working Notes of the AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments,* 1995.

This paper describes the Information Manifold, a system for browsing and querying multiple information sources, and one of the pioneering research projects in information integration. Their approach uses knowledge representation technology to organize sources, translate queries, and retrieve information.

[SL90] A. Sheth, and J. Larson. "Federated Database Systems for Managing Distributed, Heterogeneous, and Autonomous Databases". *ACM Computing Surveys*. 22(3):183-236, September 1990.

This article is representative of the work being done in the early 1990's to integrate autonomous database systems, discussing how various federated database system architectures can be developed and the critical issues related to developing and operating such a system.

[TTCB+90] G. Thomas, G. R. Thompson, C.-W. Chung, W. Barkmeyer, F. Carter, M. Templeton, S. Fox, and B. Hartman. "Heterogeneous Distributed Database Systems for Production Use". *ACM Computing Surveys*. 22(3):237-266, September 1990.

This article is representative of the work being done in the early 1990's to integrate heterogeneous database systems, outlining the approaches for various aspects of distributed database management, and describing several systems developed for production use.

[Ull97] Jeffrey D. Ullman. "Information Integration Using Logical Views". *Proceedings of the 6th International Conference on Database Theory* (ICDT'97): 19-40. Delphi, Greece, January 1997.

This paper discusses the theories behind much of the common research in information integration, specifically dealing with the concept of creating logical views that represent the capabilities of information sources and translating queries to that view into queries to the information sources.

[WN] WordNet home page: http://www.cogsci.princeton.edu/~wn/w3wn.html

This web site gives information about WordNet, including the history of the project, the scope of the database, and the database with source code libraries available for download.

[XML] The XML Catalog, listing organizations producing industry-specific XML DTDs: http://xml.org/xmlorg_registry/index.shtml

This web site lists various companies and organizations that are developing industry specific and cross-industry XML DTDs in various domains.

IX. Artifacts

The program that implements the proposed process to determine information integration mappings based on WordNet and context will be written in Java. This research will also produce a general integration framework, which will also be written in Java as well as XML DTDs for the input documents (source and target schemas) and output documents (possible and most likely mappings between target and source schemas).

X. Signatures

This proposal, by David Jackman, is accepted in its present form by the Department of Computer Science of Brigham Young University as satisfying the proposal requirement for the degree of Master of Science.

David W. Embley, Committee Chairman

Deryle Lonsdale, Committee Member

Bryan S. Morse, Committee Member

J. Kelly Flanagan, Graduate Coordinator