Human Computation Recommender for Inter-Enterprise Data Sharing and ETL Processes

Michael J. Herold, Satyajeet Raje, Jay Ramanathan, Rajiv Ramnath, and Zhe Xu
Department of Computer Science & Engineering
The Ohio State University
Columbus, Ohio 43210–1277
{herold, raje, jayram, ramnath, xuzhe}@cse.ohio-state.edu

***Abstract*—Extraction, Transforming and Loading (ETL) is an important and complex process of data integration. In this paper, we review existing ontology- based ETL methods. We then analyze why ontologies could be a better approach to ETL compared to other approaches such as folksonomies. Following this, we describe inter-enterprise collaboration, a class of problem that requires improvements to the overall ETL process. We also describe and analyze three case studies to illustrate their importance to this research.**

***Index Terms*—Expert finding, Extraction-transformation-loading, Knowledge management, Ontologies**

I. Introduction

Extraction-Transformation-Loading (ETL) tools are responsible for extracting data from candidate sources, cleaning and transforming the data to a specified staging format, and loading the staged data into a data warehouse. While general tools to aid this process exist, ETL is still done ad-hoc in many companies.

To hasten the creation of ETL processes, we want to utilize upper ontologies to aid the transformation process. However, use of an upper ontology on an arbitrary data source is not a straight-forward process. Because data representation is inherently a non-deterministic process, there are a virtually infinite number of ways to represent any given domain of data. As such, each ETL process must go through a discovery phase to create a mapping between a new data source and an established data warehouse. This is the most time-consuming aspect of ETL, and as such, is the area that we hope to improve.

There has been a lot of work on different methods to perform ETL services. These services deal not only with transformation, which is a significant part no doubt, but also integration of data from multiple sources. With the advent of the semantic web and the concept of ontologies, ETL processes have evolved to provide for not only faster, but more effective techniques to achieve their targets. This paper surveys these new ETL services and tries to evaluate their effectiveness. Our motive for doing this research is to estimate how ontology-based ETL processes provide benefits to other data-based services later on.

A. ETL as a Means to Create Innovation

Increasingly our ability to innovate, collaborate and deliver more targeted consumer services (e.g. crime prevention, early intervention, and proactive home healthcare) in key areas like health care, law enforcement, government, and academia requires us to associate, analyze and make decisions using data from multiple sources. While our industry enterprise partners agree to the mutual benefits of quickly sharing this data for more dynamic improvement of services, they are severely handicapped by the lack of needed collaboration services.

To make it easier to create inter-enterprise collaboration, there must be a method to identify experts that have specific knowledge about data and data sources, as well as potential knowledge that could be extracted to aid decision-making. To this, we claim that any non-routine need for knowledge requires the following collaboration protocol:

1) A party asks a complex question that requires information from many sources to answer.
2) Experts are found and connected to each other in order to refine the question into sub-problems, identify data sources, or identify the knowledge extraction steps necessary to determine the sub-problems.
3) Organizations (or individuals) that own the data enter into an agreement or understanding to collaborate, for instance, by sharing data or applying tools.
4) Solutions to past sub-problems are used to make future collaborations efficient.

II. Related Work

Researchers have taken many different approaches to improve the ETL process. Most techniques revolve around either a model of the data or a model of the process itself. ETL process modeling is intended to make the mapping process of one data source to the standardized data warehouse structure easier by abstracting the mapping process in various ways. Vassiliadis and Simitsis use an abstraction around the idea of concepts [1] [2]. By generalizing models of data sources from a relational model to a conceptual model, they provide a logical mapping of generalized concepts to similar concepts found in a data warehouse. This eliminates problems that arise from data source structure, whether it is from a relational database, flat files, or any other representation. It follows that,

if an automated process can form these concepts for mapping, the transformation task of ETL can be greatly hastened.

Alternative approaches to modeling the ETL process have also been researched. UML activity diagrams can be used to model the ETL process, thus grounding the process in a common language, as discussed in [3]. The advantage of this approach is that the process modeling can be done by a wide variety of people, due to the market penetration of UML.

A solid process modeling technique should be able to be standardized around a set of procedures, thus automating the generation of the ETL process from a conceptual model. Muoz et al. attempt this in [4]. By utilizing model-driven architecture and formally defined Query, View, Transformation (QVT) transformations, the authors provide a means of doing automatic code generation for ETL processes.

While process modeling can provide a robust set of techniques to reduce the cognitive load of ETL process designers, there is a notable lack in attempts to evaluate the performance of these techniques. In [5], the authors propose a set of experiments to validate measures of the performance of their UML activity diagram technique. This shows significant utility in maintainability measures, such as the number of elements in the ETL process and the number of input and output flows in the ETL process. By showing the efficacy of these measures of maintainability, the authors open the door to discovery and validation of other measures, such as ease of construction and the ability to automate the process.

When discussing ETL processes, one must take into account the wealth of research in related fields. Schema mapping, a technique that is key to information integration, is a restatement of the ETL problem. Clio, a project that creates tools for information integration, uses techniques for generating queries that represent key concepts within a data source [6]. Using the mapping to the central concept, one can match the fields of a query from one data source to the fields of a query from a different data source. This technique has potential for expansion; by integrating it with the work in [1] and [2], the mapping could automatically be generated for an ETL concept to integrate the data into the data warehouse.

Work has also been done in the area of automatic ontology matching. Given our goal of utilizing ontologies to hasten the ETL process, work in automatic ontology matching is an important avenue for investigation. In [7], Mascardi et. al. evaluate the use of upper ontologies for automatic ontology matching. The authors evaluated their techniques using several upper ontologies, including SUMO-OWL, OpenCyc, and DOLCE. This work will be important to keep in mind as we develop our technique.

III. Method

A. Why Use Ontology-Based ETL

Ontologies have been used in a variety of biomedical applications, such as data integration. This approach to integration usually involves annotating multiple sources of data using some controlled vocabularies. This method has proven to be very successful. An excellent example would be the Gene Ontology which has been used efficiently to standardize genes and gene products using a controlled vocabulary. The efficacy of this method has led to several domain specific and ad-hoc ontologies. The Open Biomedical Ontologies (OBO) [8], which includes the Gene Ontology, is a consortium which tries to keep track of and coordinate these efforts.

The most interesting aspect is that the data that is stored using ontologies is very well structured. There are several tools, such as AmiGO for the Gene Ontology [9], which can perform data mining tasks very effectively. An obvious question is whether similar techniques would perform as well for domains other than biomedicine. Also, we must look at ETL using ontologies as a precursor to developing such data-based applications. This might be one of the biggest benefits provided by ontology-based ETL practices. The example of the Gene Ontology indicates that once the data is integrated in a structured manner along with its conformation to an accepted ontology, it can be used to provide a layer of services to users which would have been difficult to otherwise achieve.

In the rest of this section, we first compare ontologies with folksonomies. Secondly, we compare and analyze different types of ontologies. Finally, we analyze some representative ETL methods.

B. Ontologies vs Folksonomies

Ontologies, as shared understandings of concepts and the relationships between them, are usually developed by domain experts. During the development, knowledge engineering coordinate term usage and achieve agreement on a set of terminology. Ontologies are beneficial since they provide high consistency and accuracy. However, they introduces the issue of scaling. It is hard to ask for complete agreement on terminology with a large ontology. Another disadvantage is that ontologies lack the ability to deal with polymorphism. For example, a term may have several different meanings and several different terms may refer to the same concepts.

Folksonomies are another approach that complements ontologies. Human computation and crowdsourcing are utilized to generate folksonomies. One example of a folksonomy is the social bookmarking service “Delicious”, which consists of user-tagged bookmarks. Folksonomies provide a more flexible way to achieve large scale knowledge sharing, which is costly to accomplish using ontologies. However, high flexibility also makes folksonomies fail to achieve high consistency and accuracy. Considering accuracy is one of the focuses of ETL, ontologies are more appropriate than folksonomies to support ETL process.

C. Upper, Middle, and Lower Ontologies

As mentioned earlier, ontologies are costly when attempting full domain coverage due to scaling issues. Thus, we classify ontologies into three different types: upper ontologies, middle ontologies and lower ontologies. We believe a good use of this classification can help solve the issue of scale.

Upper ontologies are sharable understandings of concepts which are applicable across domains. Serving at a top level, upper ontologies are the first step for an ETL developer to consider. Starting from an upper ontology could improve

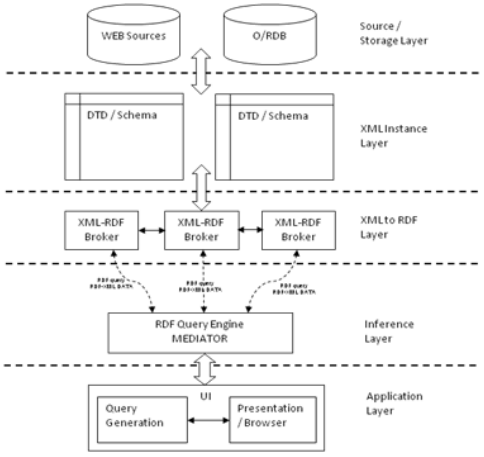

Fig. 1. A model of semantic integration.

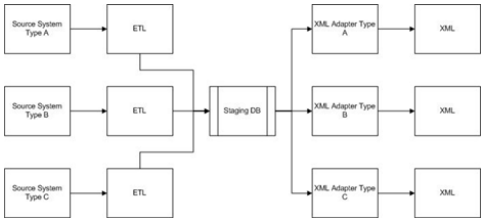

Fig. 2. A characterization of the Sypherlink ETL problem.

efficiency, since the developers don't have to start from scratch. Lower ontologies are specifications of particular domain. Middle ontologies serve as bridges connecting various upper and lower ontologies.

A good approach is to look through ontologies in a top-down fashion, from upper to middle to lower ontologies. This is top-down approach is based on the intuition that the focus of ETL processes is consistency instead of individual differences.

D. Ontology-based ETL Methods

Semantic web [7], or Web 3.0, has opened new avenues in cyberinfrastructures. It has allowed the web itself to be more "structured". Data extraction over the Internet has become a well-trod practice for many enterprises dealing in knowledge management. Data integration and, more specifically, "semantic" integration have become mainstream techniques for acquiring data.

The basis of OWL-based ETL is making use of the semantic metadata which is an integral part of Web 3.0. OWL- or RDF-based ETL is becoming very popular. One of the major reasons for this could be that OWL is now a recommended standard for representing ontologies according to the W3C. Since it is supported by XML it is very easy to integrate and use.

Figure 1 is a framework based on semantic integration.

IV. Case Study

A. Sypherlink

Sypherlink, a software firm based in Dublin, Ohio, is dealing with the process that integrates different data sources based on NIEM ontology, as illustrated in Figure 2. The data mapping

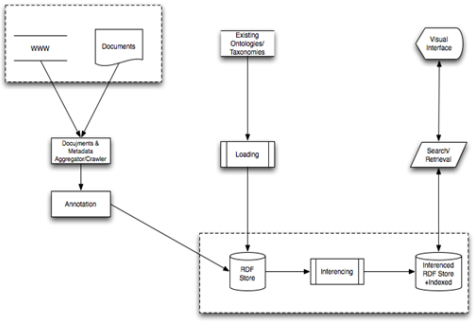

Fig. 3. A system overview for the project at The Ohio State Medical Center.

stage takes only several weeks while the data moving stage takes months. A lean strategy is to focus on the data moving first, which would return high business value. We are still on the track of figuring out which technique would be optimal for this Sypherlink problem.

B. Biomedical and Healthcare Applications

An ongoing project initiated by the Ohio State University Medical Center is exploring the possibility of using ontologies from the OBO consortium to develop a search tool. The goal is multifold: build a semantically- anchored search engine that could seamlessly query across all of these resources and platforms; provide workflow support that will allow for the seamless integration and pipelining of the available resources and architectures; and both automated and semi-automated data retrieval with the help of ontologies. This project will provide insights on the effectiveness of this technique and help address the above issues.

The project uses the OWL/RDF framework to store the annotated resources. SPARQL, in conjunction with graph-based data mining techniques, is used to query the model and rank the obtained results. An overview of the system can be seen in Figure 3.

C. Utilizing ETL to Identify Points of Collaboration

To illustrate, many development experts in the City of Columbus used knowledge of multiple diverse data sources: 1) demographic information; 2) local fast food locations and size; 3) incidence of type II diabetes; 4) tax data; 5) location of grocery stores; and 6) connections with the local university policy makers and experts. This investigation took considerable time and many meetings but eventually resulted in the formulation of a tax incentive to a specific fresh food cooperative to establish a new location in the underserved area. This type of collaboration between policymakers, entrepreneurs and researchers through the exchange of data and expertise from their respective enterprises for mutual value is the eventual goal of our proposed prototype system.

V. Evaluation

The evaluation is to be done using three case studies from different domains. The first is the Sypherlink project which explores the application domain of e-governance and uses

the National Information Exchange Model. The second is the Ohio State Medical Center initiative to develop a semantically rooted search engine for health care researchers which makes use of ontologies from the Open Biological Ontology discussed above. Using these two studies we should be able to answer both the questions presented by ontology-based ETL. These are "Is there an improvement in the performance of the actual ETL process?" and "Will using ontologies to do ETL have a benefit subsequently in the form of data mining or data warehousing services?"

The third case study is the experience we noted in the collaboration with the City of Columbus to identify policies to improve the health of under-served diabetics. This case study revolved around the collaborative process of sharing data and expertise. Complexity, heterogeneity, noise, the sheer mass of data, a lack of the right knowledge, and the inherent difficulty of predicting changing behavior all combine to make data sharing unreliable. To facilitate this, we will develop collaboration protocols for data sharing that complement existing machine algorithms, such as ETL and query filters applied across structured and unstructured, semantically rich content, by incorporating collaborative recommendations for "human computation" by experts. These algorithms will be designed here to utilize the effective and efficient connection discovery that machines do well in order to leverage human problem-solving abilities for problems that machines cannot handle.

This work will seek answers to the following questions: How can we use computational algorithms to support collaboration protocols that result in the connections between humans with appropriate expertise? Can human computation be leveraged to make the inferences needed to more effectively share data? How can we recommend, in a timely manner, person- to-person connections for collaboration with experts? What motivates organizations to collaborate and how can we engage individuals within those organization?

For this research we will initially use two data sets: one from the National Ocean Council and one from the Department of Energy. We will later scale this work to law enforcement with one of our industry partners.

The intended outcomes of this research include: 1) an analysis of the feasibility of a recommender algorithm for a data collaboration protocol; 2) a case study that will illustrate the recommender's effectiveness in sharing data, especially in contrast to existing tools and methods; 3) a demonstrable prototype to engage potential users; and 4) a framework for further research.

VI. Conclusion

In this paper, we proposed a way to analyze different ontologies by grouping ontologies into a hierarchy of upper, middle and lower ontologies. The demand for ETL and data integration especially from unstructured data sources like the web is already high and increasing at a steady pace. The objective of this research is to check if ontology- based ETL processes provide a novel and productive solution to this problem. We also discuss the potential use of these methods to aid in expert finding for collaboration. This type of task is vital to increase inter-enterprise innovation.

References

[1] P. Vassiliadis, A. Simitsis, and S. Skiadopoulos, "Conceptual modeling for etl processes," in *Proceedings of the 5th ACM international workshop on Data Warehousing and OLAP*, ser. DOLAP '02. New York, NY, USA: ACM, 2002, pp. 14–21. [Online]. Available: http://doi.acm.org/10.1145/583890.583893

[2] A. Simitsis and P. Vassiliadis, "A methodology for the conceptual modeling of etl processes," in *In Proc. of DSE'03*, 2002.

[3] L. Muñoz, J.-N. Mazón, J. Pardillo, and J. Trujillo, "Modelling etl processes of data warehouses with uml activity diagrams," in *Proceedings of the OTM Confederated International Workshops and Posters on On the Move to Meaningful Internet Systems: 2008 Workshops: ADI, AWeSoMe, COMBEK, EI2N, IWSSA, MONET, OnToContent + QSI, ORM, PerSys, RDDS, SEMELS, and SWWS*, ser. OTM '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 44–53.

[4] L. Muñoz, J.-N. Mazón, and J. Trujillo, "Automatic generation of etl processes from conceptual models," in *Proceedings of the ACM twelfth international workshop on Data warehousing and OLAP*, ser. DOLAP '09. New York, NY, USA: ACM, 2009, pp. 33–40. [Online]. Available: http://doi.acm.org/10.1145/1651291.1651298

[5] ——, "A family of experiments to validate measures for uml activity diagrams of etl processes in data warehouses," *Inf. Softw. Technol.*, vol. 52, no. 11, pp. 1188–1203, Nov. 2010. [Online]. Available: http://dx.doi.org/10.1016/j.infsof.2010.06.003

[6] R. Fagin, L. M. Haas, M. Hernández, R. J. Miller, L. Popa, and Y. Velegrakis, "Conceptual modeling: Foundations and applications," A. T. Borgida, V. K. Chaudhri, P. Giorgini, and E. S. Yu, Eds. Berlin, Heidelberg: Springer-Verlag, 2009, ch. Clio: Schema Mapping Creation and Data Exchange, pp. 198–236.

[7] D. Allemang and J. Hendler, *Semantic Web for the Working Ontologist: Effective Modeling in RDFS and OWL*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[8] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, and S. Lewis, "The obo foundry: coordinated evolution of ontologies to support biomedical data integration," *Nat Biotech*, vol. 25, no. 11, pp. 1251–1255, Nov. 2007. [Online]. Available: http://dx.doi.org/10.1038/nbt1346

[9] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, "Gene ontology: tool for the unification of biology. the gene ontology consortium." *Nature genetics*, vol. 25, no. 1, pp. 25–9, May 2000.