Semantic Analysis of Public Health Information

Brad Schell and Kasi Periyasamy
Computer Science Department
University of Wisconsin-La Crosse
La Crosse, WI 54601
{schell.brad, kperiyasamy}@uwlax.edu

Abstract

Semantic data analysis involves interpreting data using their relationships and their purpose in real-world applications. While a simple database provides data that can be interpreted by humans, semantic data analysis requires the semantics (relationships between data) also to be stored and retrieved so that computers can interpret the data. Semantic data models have been extensively studied for at least two decades. Recently, semantic web technologies have emerged as extended versions of semantic data models which leverage web technologies. In this paper, the authors describe an ongoing project in developing a tool for semantic data analysis for the data collected by the Public Health Department at the University of Wisconsin-La Crosse. A prototype has been developed using a semantic model that represents a portion of the data being collected. Ongoing work is discussed in detail in the paper.

1 Introduction

As part of an on-going community project, the Public Health Department at the University of Wisconsin-La Crosse collects data about family nutrition, family health, medical data and literacy data from various communities, analyzes them from different perspectives and provides appropriate advices/suggestions to those communities in order to improve their health. The data collection process is usually done through surveys. Some of them are structured in Excel format while others are free-form data with no specific structures. It is therefore difficult for the department to analyze the data and make appropriate decisions. In this context, the department decided to use a software system that will assist the data collection and analysis processes. This paper focuses on the development of the software system that will assist the department in the analysis of data.

Data analysis as described above is similar to data mining. The latter is usually performed on a huge collection of data which may have several dependency relationships among the data. A key area of analysis in this case is semantic data analysis in which the relationships between data items are expressed as semantic dependencies. For example, the data collection regarding family relationships may include people of several generations apart. The analysis process will answer to typical queries such as how many cousins a particular person has. A more detailed analysis may reveal subtle information such as how many descendants in the same family have a particular disease. The challenge for this type of analysis lies in choosing the appropriate data structures so that the queries on the data can be performed efficiently and correctly.

This paper describes the design and implementation of a prototype on family hierarchies. The extension of this prototype is currently in progress towards the design and implementation of the data analysis tool for the Public Health Department.

2 Semantic Data Analysis

2.1 Semantic Data Model

A semantic data model describes data entities along with the relationships among them so that the meaning of the stored data can be interpreted by retrieving the relationships along with the data. The notion of semantic data model has been studied for more than two decades. A detailed discussion on semantic relationships and their importance in semantic data models can be found in [3, 8]. A schematic of a semantic data model is shown in Figure 1. The semantics of data is derived from the application domain and are expressed generally as constraints in the conceptual model. Thus, while storing and retrieving data from the physical layer, the database system enforces the constraints (semantics of the application domain) on the data. Usually a separate semantic data model is built for each application. For example, Hayashi and others [4] described a semantic model for medical information management system that assists in diagnosis of dementia. The semantics of the data in their model come from the medical data files already developed by the re-

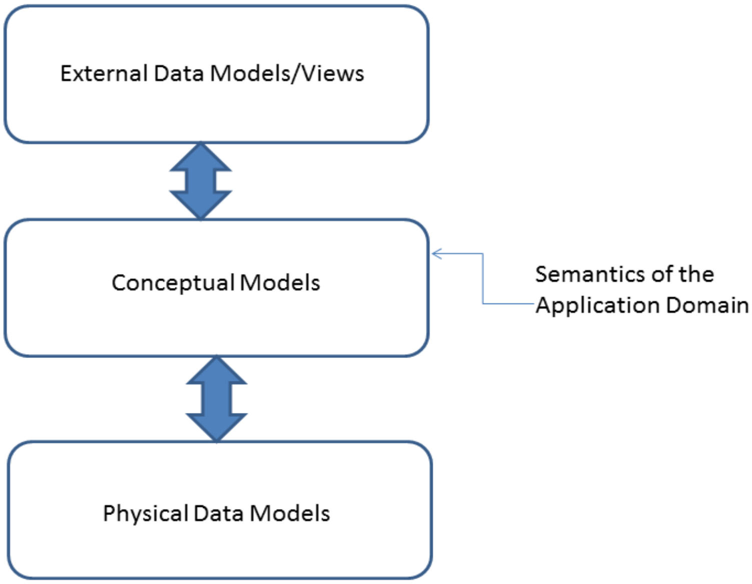
External Data Models/Views
Conceptual Models
Semantics of the Application Domain
Physical Data Models

Figure 1: A Schematic of Semantic Data Model.

searchers. This paper is a contribution to the design of a semantic data model for public health information.

2.2 Semantic Web

Recently, semantic web technologies have become popular in data analysis. These technologies are extended versions of semantic data models described earlier but focus on web-based data; that is, data passed through the web. A gentle introduction to the semantic web is given in [1]. Both semantic data model and semantic web technologies have one thing in common: the meaning of data must be stored along with the data so that data processing units could interpret them correctly in order to take appropriate actions. While semantic data models are more specific towards certain applications, semantic web technologies use extensive data and complex relationships and also utilize the web technologies. As a result, the processing and interpretation of data reach a wider audience than a similar result in semantic data models. Recently, there are efforts in combining semantic data models and semantic web technologies together because of the similarity in their goals-attaching semantics to data [2, 5]. Figure 2 shows a stack of activities in a semantic web. Some of the key activities in a semantic web (as shown in the stack) include (1) XML markup language for syntax of data items, (2) data interchange through Resource Description Framework (RDF) notation, (3) query processing through Web Ontology Languages (OWL) and (4) additional elements to verify the correctness of data processing. Depending on the ap-

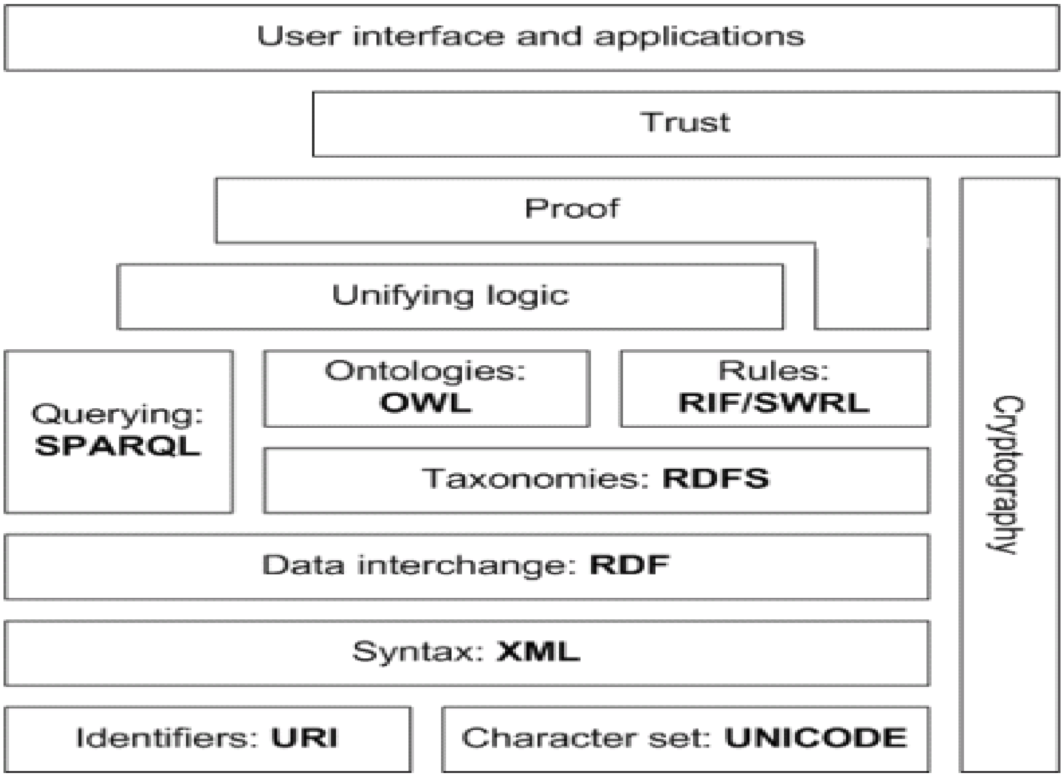

Figure 2: Semantic Web Stack©Tim Berners-Lee, 2001.

plication domain (and hence the semantics of the data), the selection of ontology for the particular semantic web becomes more complex. For example, Viswanathan and others [9] described some of the challenges they faced when designing a semantic web system for networked applications. The challenge in their application comes from the various network topologies used. Hayashi and others [4] reported similar challenges from medical applications.

3 The Current Project

The Public Health Department at the University of Wisconsin-La Crosse is gathering data from various communities in Nicaragua with the intention to provide adequate public health information that would improve their health status. Four sets of data are collected in this process - medical data, family and community health data, nutrition data and literacy data. The data collection process uses a set of survey questions and record the results of the survey in plain text and spreadsheet formats. Since the surveys include questions from different perspectives and are not geared towards an organized set of data, storing the data in a structured format is almost impossible, especially during the data collection process. As a result, the data analysis has to be deferred until the data collection process is completed. Moreover, the remoteness of the site also plays a crucial role and stumbling block to provide faster feedback to the participants. A typical cycle of data collection, data analysis and feedback takes almost a year. A sample set of survey questions for family nutrition is shown in Appendix.

In order to expedite the feedback and to gather more, fine-tuned data, the Public Health Department needs some assistance in faster data processing and assessment. The current project is aimed at developing an interactive software system which would process data on the site and provide as much feedback as possible. Though it is not possible to provide feedback on every single data item immediately, at least some major and important information can be provided to the participants before the next cycle. This project is the first step in developing a data analysis tool that supports the team that conducts the surveys.

The Public Health Department team is interested in performing a qualitative analysis of data rather than quantitative analysis; the latter can be easily performed using a simple database. This requirement gave the motivation for us to develop a semantic model rather than simple numeric analysis of data. The assessment of data collected in this project not only depends on the surveys but also relies on the comparison of this set of data with published results. For example, the nutrition data collected from a set of families may indicate that these families lack a particular diet in their meals but this information can only be confirmed if the nutrition data is compared against the standards published. While it is possible to store some of this information in a database, the huge volume of data to be processed and their complex relationships warrant an on-line search of this information. Thus, a semantic web technology might be appropriate choice for this software system. Currently, we just completed a semantic data model for a portion of this application in order to gain understanding of the application domain and the complexity of relationships among data. A prototype implementation of the semantic data model has been described in the next section. We are investigating the semantic web technologies to use them in the next phase of our work.

3.1 Prototype - semantic data model

We started with a semantic data model of a portion of the health care application. This model describes family members and their relationships. A family hierarchy includes people who are typically several generations apart. For simplicity, the names of the people in one family hierarchy are assumed to be unique. While it is quite common for more than one person in the same family hierarchy to have the same name, it is always possible to augment such duplicate names with some additional information such as suffixes (Jr., Sr. etc.) or some numbers indicating generation level or birth year. Figure 3 shows a typical family hierarchy. It lists people with connections to their immediate family members (father, mother, siblings and children). To simplify, only immediate relationships are included in the diagram. Relationships such as step-father and step-mother are not included in this model. The names on the leftmost end indicate the set of people selected for this prototype. The next level indicates the children of the couple on the leftmost end. For example, Jim and Robin have three children named Alan, Sarah and Nancy. Sarah and Roger have three children named Joe, Bill and Britney. The relationships in this diagram (the semantics of the domain) are the immediate blood relationships among the family members. From left to right, the relation indicates immediate descendants (sons and daughters) while from right

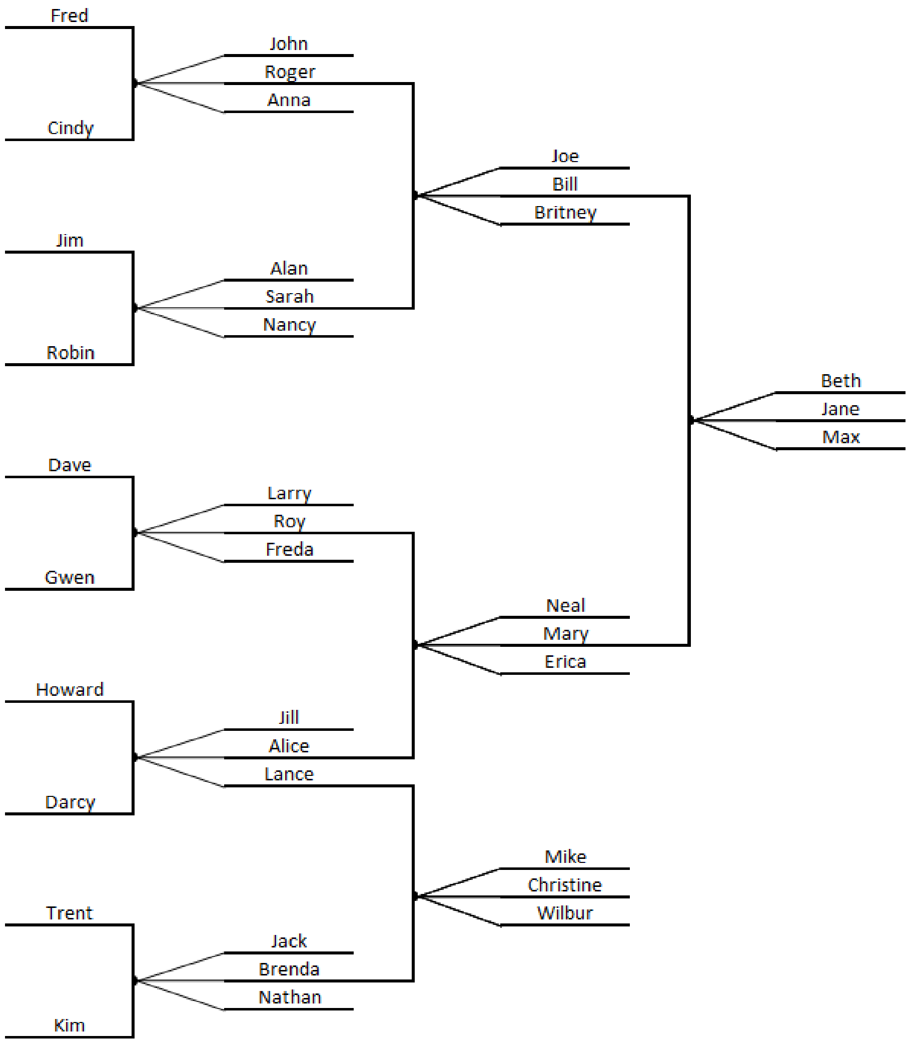

Figure 3: A family hierarchy.

to left, the relation shows parents.

The query processing for this application includes (1) finding the descendant relationships such as how many children or grandchildren a particular person has, (2) finding ancestor relationships such as who are all the ancestors for a particular person, and (3) showing the relationship between two people. For example, one might ask the question “how Wilbur and Howard” are related. In this case, the trace shows that Wilbur is related to Lance (parent) and Lance in turn is related to Howard (parent), thus Howard is a grandparent of Wilbur.

As stated earlier, the semantics of the application in this case is stored in the links between names.

4 Conclusion and Continuing Work

This paper describes an ongoing work in developing an interactive system that analyzes public health data. The analysis focuses on providing educational support to a rural community in Nicaragua in terms of nutrition, public and family health system. The research team will be using this tool on site during data collection in order to provide first hand feedback to the participants. The software development team has developed a prototype using a semantic model for a portion of this problem. Our immediate next step is to enrich the semantic data model to include additional relationships and trace them through query processing. We are also investigating semantic web technologies and data mining technologies [7].

Acknowledgement

The authors would like to acknowledge Dr. Gary Gilmore of the Public Health Department at the University of Wisconsin-La Crosse for sponsoring the project and providing data for the research.

References

[1] Berners-Lee, T., Hendler, J. and Lassila, O. 2001. The Semantic Web. *Scientific American*, 284(5), 34-43.

[2] Bussler, C. 2008. Is Semantic Web Technology Taking the Wrong Turn? *IEEE Internet Computing*, Jan/Feb (2008), 75-79.

[3] Chang, T. and Sciore, E. 1992. A Universal Relation Data Model with Semantic Abstractions. *IEEE Transactions on Knowledge and Data Engineering*, 4(1)(1992), 23-33.

[4] Hayashi, M. *et al.* 2008. A Medical Information Management System Using the Semantic Web Technology. In *Proceedings of the Fourth International Conference on Networked Computing and Advanced Information Management (NCM 2008)*, 2, IEEE Computer Society Press, 75-80.

[5] Janev, V. and Vraneš, S. 2009. Semantic Web Technologies: Ready for Adoption? *IEEE IT Pro*, Sep/Oct (2009), 8-16.

[6] Pham, D.T., Dimov, S.S. and Huneiti, A.M. 2003. Semantic Data Model for Product Support Systems. In *Proceedings of the IEEE International Conference on Industrial Informatics (INDIN 2003)*, Alberta, Canada, 279-285.

[7] Rubin, S.H. 2013. On the Associative Mining of Big Data for Creating Decision Making. In *Proceedings of the International Conference on Computers and Their Applications (CATA 2013)*, Honolulu, Hawaii, 2013.

[8] Storey, V.C. 1993. Understanding Semantic Relationships. *VLDB Journal*, 2(1993), 455-488.

[9] Viswanathan, A. *et al.* A Semantic Framework for Data Analysis in Networked Systems. 2011. In *Proceedings of the 8^{th} USENIX conference on Networked Systems Design and Implementation (NSDI'11)*, USENIX Association Berkeley, CA, USA, 2011.

Appendix

Sample questions on family nutrition used during the interviews:

- How many meals per week are eaten away from home?
- What does a child in this family usually drink? (choices given)
- Does anyone in the household not eat some foods because of health reasons? If so, what types of food?
- List where you get your food (e.g., local market) and how often?
- Who commonly prepares food for the family?
- What do you do with the remaining food, if food remains after meals?
- What foods would you and your family like to eat more often?