2014 IEEE 14th International Conference on Bioinformatics and Bioengineering

An Ontology-based Framework for Analysis Recommendation

Gabriela Henriques
School of Computer Science
University of Guelph
Guelph, Canada
ghenriqu@uoguelph.ca

Deborah Stacey
School of Computer Science
University of Guelph
Guelph, Canada
dastacey@uoguelph.ca

***Abstract*—One of the challenges with data analysis revolves around selecting the best analysis method for a data set that will provide appropriate and meaningful results. This paper presents an ontology-based framework to address challenges around selecting an analysis method that can best represent a data set and the information you want to get out of it. Two ontologies were developed, one to capture semantic and syntactic descriptions on a data source, and likewise one to capture the description of analysis methods. Ontologies were selected for their flexibility in providing a description between a set of concepts and relationships along with their ability to reason between these descriptions.**

***Keywords*-Ontology, Knowledge Engineering, Data Analysis, Analysis Recommendation, Leveraging Knowledge, Syndromic Surveillance**

I. Introduction

As methods for data collection increase, the need for better data analysis also increases. In recent years we have seen a rise in techniques and technologies specializing in collecting data for a domain. However, data analysis is not improving in such a rapid pace; there are many factors that goes into selecting the best analysis method on a set of collected data in order to attain meaningful and useful results [8]. Many systems specialize in analyzing data and providing visualization to an end-user, but often times the challenge lies with how to interpret a set of data. People are constantly trying to figure out what their data means, or what their data is telling them. In order to do so, data is first analyzed and then interpreted. However, if an incorrect form of data analysis is used on a set of data, this will skew how it is interpreted and can provide meaningless or incorrect results.

This paper discusses an ontological approach for describing a data source and determining the most appropriate data analysis method for a set of data, focusing on the domain of syndromic surveillance. The early stages of a framework was proposed in the paper *An Ontology-based Framework for Syndromic Surveillance* [7]. This paper expands on the initial research, focusing on ontology design for leveraging knowledge.

II. Background

A. Challenges with Analysis and Method Selection

At times, there are various sources from which data is collected for analysis. In regards to syndromic surveillance, some of the forms of data include emergency department diagnostics, over-the-counter pharmaceutical sales, and news reports [1], [3], [9]. In this domain, data can be grouped into three different categories of sources: pre-clinical, clinical pre-diagnostic and diagnostic [1]. In order for analysis methods to be accurate and effective, it is important that the data is specific. However, due to the need for timely detection, it has become more popular to analyze clinical pre-diagnostic information [2]. This data incurs a loss of specificity during the collection process, some of the algorithms used for detection may be ineffective without taking extra precautions on how to interpret the data. This describes the need for having a description of the source(s) of data when going through analysis.

Systems are composed of many types of users ranging from novice to experts. In the case of syndromic surveillance, a typical end-user consists of a health analyst or epidemiologist who analyzes a data set and determines whether a disease outbreak is occurring. There are various different methods which can be used for conducting this analysis; in order to determine which method is best suited for a set of data, the user performing the analysis would have to be an expert in all systems and methods. This is usually not the case; for example, the user may be knowledgeable in various statistical methods, which are commonly used for analysis, but may not consider other statistical methods that can provide more value for the analysis that they need. This can also be seen when combining methods from other fields to use for data such as neural networks or genetic algorithms. Often times, the end user will not have a sufficient background in the field of computer science to understand these algorithms or the type of results they can provide through analysis.

Determining how to analyze a data set is also influenced by the type of information a user is aiming to discover

978-1-4799-7502-0/14 $31.00 © 2014 IEEE
DOI 10.1109/BIBE.2014.70

IEEE computer society

from their data (i.e. their goal), as well as other factors of importance to the end user such as trust, performance or quality. Their goal and factors can be defined differently for every set of data; this should be taken into consideration when the user is selecting the most appropriate method(s) to meet their needs.

B. Leveraging Knowledge

Leveraging knowledge encompasses the concept that the transfer of knowledge between two people is bi-directional and that knowledge grows when used and depreciates when unused [4]. This is an important concept to consider when discussing challenges with data analysis as it can be incorporated to help address some of the challenges described previously. If we look at a domain expert such as an epidemiologist, whose job is to study cause and patterns of diseases and outbreaks, the epidemiologist may be knowledgeable in methods typically used in their domain. However, a computer scientist can also be knowledgeable of algorithms that can be applied in order to discover patterns in a set of data. Since these algorithms are not typically used in the domain that the epidemiologist was trained in, they may not have sufficient knowledge or technical background to consider the algorithms that the computer scientist can consult on. Providing a description of analysis methods in a manner that is understood by users regardless of domain or expertise level can help eliminate the interpretation barrier explained in the above example.

C. Ontologies

An ontology is a form of knowledge representation; it is used to represent a set of concepts and their relationships within a domain [6]. An ontology has the ability to reason with the entities of a domain and can thus be used to describe the domain itself [6]. Many ontology-based frameworks have been developed in various application areas to aid in data collection, organization, and classification.

Ontologies have been selected in this research for their ability to syntactically and semantically describe data. With this ability, we can accommodate leveraging of knowledge between multiple domains. It is important to have a description on the data that has been collected; this can aid in the analysis process by helping us select the best mechanism for analysis based on the data set. Along with the need of having better descriptions of data, we identified the need for an end user from any background and with any level of knowledge to be able to understand what an analysis method can provide with their data set. Selecting the correct analysis method is an important step in the process of data analysis; in syndromic surveillance the results attained and the meaning of those results will depend on the analysis method used and could identify whether an outbreak is occurring.

III. The Framework

The purpose of the ontology-based framework is to provide an end user with a set of techniques that they can apply to their data set based on the type of analysis they want to perform on the data. The framework aims at validating analysis techniques that a user would want to use on a set of data, as well as providing a recommendation of methods that they may use and what those methods will bring to them through the use of ontologies. As well, for the system to be effective it must be able to eliminate any barriers formed between domain experts and incorporate this notion of bi-directional knowledge sharing. The ontologies used within the framework will reason based on a set of parameters provided by the user.

A. System Architecture

The proposed framework is composed of 4 layers: View, Application, Knowledge and Source. The View layer contains the interface; what an end-user can see. This layer involves gathering input from the user and displaying the output. The input would typically involve the set of data that the user would like to analyze, along with the analysis goal they wish to achieve in the form of a question. The Application layer is composed of the system that connects with the ontologies and the interface and is responsible for the interaction between the two. The Knowledge layer contains two ontologies which are used together in the process of selecting the analysis method best suited for the set of data. Lastly, the Source layer provides the various sources for the data that is going through analysis, along with any sources that can be used to connect with an analysis method including systems that the method can be run on. Figure 1 provides an example of what a fully implemented application architecture would be like using the proposed framework.

The framework concentrates on providing an Application and Knowledge layer that can be adapted to any domain. The View layer is dependent on the needs of the user while the Source layer can change when dealing with different data sets and different domains. The Application and Knowledge layers are flexible in that they have been developed abstractly enough to allow for use within other systems. It is also plausible to use the Knowledge layer independently, thus providing strictly the ontologies themselves that can be used on any type of data. The Application layer depends on the Knowledge layer as it is working closely with the ontologies. When changing the structure of the data source or algorithm ontology, the Application layer will need to be updated to accommodate

the structural changes made.

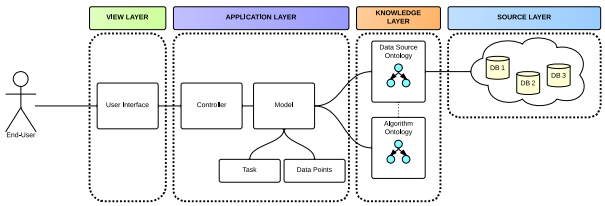

Figure 1. System architecture of proposed framework

B. The Knowledge Layer

The Knowledge layer is an essential part of the proposed framework; it contains the portion of the framework that is responsible for leveraging knowledge between multiple domains. The Knowledge layer is composed of two ontologies: the data source ontology (DSO) and the analysis methods ontology (AMO).

The DSO aims at capturing the semantic and syntactic descriptions of a data source; a data source typically contains one or more data sets, each set containing one or more data points. The AMO aims at describing semantic and syntactic information about an analysis method (statistical, algorithmic or other).

IV. Ontology Design

A. Data Source Ontology

As described in section 2.1.1, data collected for analysis can sometimes come from more than one source. The purpose of the data source ontology is to describe any set of data by capturing the syntactic and semantic descriptions of the individual data points. By describing each portion of the data set we come closer to describing the source of data itself and capturing semantic properties of the source [5], [13].

Before designing this ontology, research was done to determine whether an existing ontology could meet the need for describing a data set. The ontology-based application BioSTORM contains a data source ontology to describe the sources of data used in the application [11]. This ontology was not selected for use in this framework, as it provides mostly low-level syntactical description of a data source such as the data format for each data point. It would therefore not contribute sufficiently to meeting the overall need of capturing syntactic and semantic description of a data set. During further research, it was noticed that many of the ontologies integrating into applications were specific to a particular domain or designed in such a way that did not allow for adaptability into other fields or applications [11] [12] [10].

This ontology is composed of 7 main classes: Data Source, Data Type, Data Size, Data Properties, Data Points, Data Format and Data Storage.

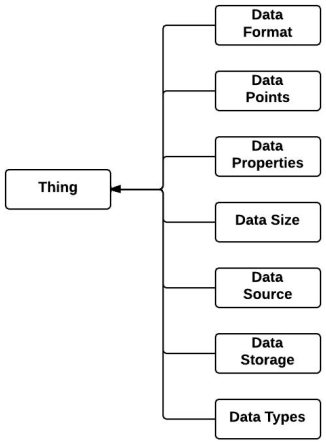

Figure 2. Main Classes of the data source ontology

The data source class contains a list of the data sources for the set of data you want to describe. This class is typically composed of one or more sources; each data source contains a set of data points and may contain a description of the size of your data set and how your data is stored. It is recommended to include an annotation with a description of the source of data. This class can act as a repository of sources for all of your data points. Figure 3 displays the data source set-up for the domain of syndromic surveillance; the classes that have been included represent common data sources in this domain.

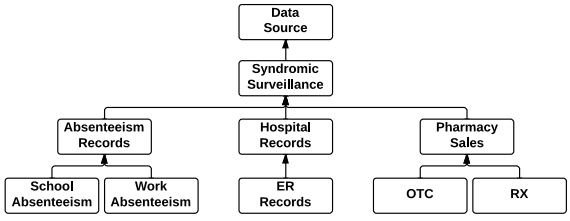

Figure 3. Data Source class set-up for Syndromic Surveillance sources

The data points class contains a list of all of the data points involved across all of the data sources that are being worked with. Each data point is described through

format, property, source, and type. We recognized that it is important to capture details on the format of each data point in the source (whether it is a string, time interval, etc.). However, we did not want this ontology to solely describe the low-level details such as the format that each data point encompasses. This would result in solely a syntactic description of each data point. Instead we wanted to allow the user to capture some semantic meaning on each data point as well. The data points contain further description through the assigning of other properties such as a data type and data properties.

All other classes are used to further describe the properties of a data source or data point and provide more contextual information.

B. Analysis Methods Ontology

Ontologies were selected as the approach to take with capturing the description of an analysis method. Technical details and meta-data on each method can be captured through a combination of classes, properties and annotations. This will allow for the syntactic and semantic descriptions of the methods. This design path also allows for the adoption of the concept of leveraging knowledge; using the ontology to act as the mediator between an end user and a background-expert (*i.e.* computer scientist, statistician). It also promotes the leveraging of knowledge between a data set and analysis methods; the semantic and syntactic descriptions of the analysis methods can be used to validate if they are compatible with a set of data points.

The ontology is composed of 10 main classes (see Figure 4). The ontology aims at capturing analysis method data needs, characteristics, and relationships to other methods.

The Methods class contains of a list of all methods that are described through object property relationships between the other main classes. The analysis types class provides sub-classes that are used to categorize the type of analysis that a method does. This class can be used to help determine what the goal of a user is in terms of how they would like to analyze their data and what type of information they are looking for.

1) Data Needs Description: Describing what is needed from the data will enable the ontology to validate whether a set of data points are compatible with an analysis method. The following classes are used to help describe the needs that a method has for data it analyzes. It is also important to note that the following classes overlap with classes found in the Data Source ontology; Data Set format, properties, size and type. In the Data Source ontology, these classes are used to describe a data point; in the analysis methods

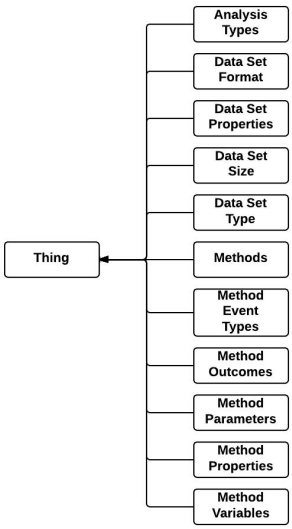

Figure 4. Main classes of the analysis methods ontology

ontology, these classes are used to help describe the input requirements of a method. A combination of the following classes will be used to describe the characteristics of method inputs, to help determine if a set of data points can be used with the method.

2) Characteristics Description: Along with describing the input requirements for an analysis method, this ontology also provides a means for describing the characteristics of a method. Providing this level of description is important as it gives further context to the user about the method, and also provides the user with the ability to go beyond input requirements when determining whether a data set can or should be used with a particular method. Classes have been included to describe the following characteristic aspects of a method: event type, outcome type, parameters and method properties.

3) Relationship Description: The object properties are used to describe the relationships between analysis methods. By describing the relationship between methods we can better determine alternative analysis methods a user may want to consider. Three object properties have been included to describe the relations between methods: hasRelatedMethod, isUsedOn, isUsedWith.

Figure 5 shows the description of the analysis method 'Mean' through the analysis methods ontology.

V. Discussion

A command-line prototype of the application layer of the framework was created to demonstrate the use of the

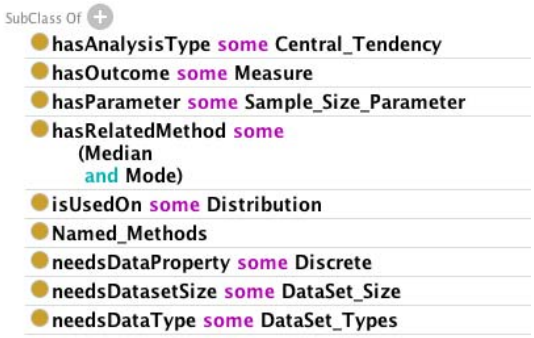

Figure 5. Sample description of the method 'Mean'

knowledge layer framework. The prototype was also used in the validation of the data source and analysis methods ontologies. A set of use-cases were developed during the validation process using Over-the-counter and prescription drug sale data provided by the Public Health Agency of Canada.

A. Data Source Ontology

This design provides a flexible ontology that allows for the description of any data source without the restriction of specifying a particular domain or type of source. This ontology also allows for the description of multiple data sources simultaneously, thus providing the ability of combining data points from various sources within the same ontology. Another advantage brought upon through the design of the ontology is that it promotes portability. This ontology can be used within any application that contains the need for data description. Since the ontology is not domain-specific, it has a higher chance of being adaptable to the needs of a user. This is in contrast to other data source ontologies such as the BioSTORM data source ontology, which was developed around the purpose of describing syndromic surveillance data. As well, the structure of the data source ontology developed by BioSTORM provides more of a syntactic description of data points through concentrating on describing the data format of each data point. Our ontology however aims at providing the data format along with other descriptions through classes. The ontology also aims at capturing semantic information on the data source through accompanying object properties as well as through describing the data points that make up that source. By doing so, this allows for the opportunity to use other ontologies to aid in the description of a data source or data point; this is seen through the example of a Data Type class where the time ontology was imported to describe data points of the type time.

B. Analysis Methods Ontology

The analysis methods ontology provides a good tool for describing the features of a method, along with information about the data that each method requires in order to run. On its own, the ontology aims at capturing the key information that a user would need to know in order to use a method. It also provides flexibility in that it does not limit the type of methods, or the type of analysis to a specific domain or type. For example, an analysis method can be defined as a tool that helps the user investigate a set of data; this could mean a statistical method, an algorithm, a system or some other form of analysis. Combined with the Data Source ontology, this ontology can be used to evaluate whether a set of data points can be used on an analysis method based on key information such as format, property(s) and data type(s).

The description of the analysis methods currently contained in the ontology have been described based on the R statistics description as a starting point. Not a lot of research has been done focused on ontologies for describing analysis methods, this ontology provides a starting point for these descriptions with the potential of expansion.

VI. Conclusion

A. Leveraging Knowledge through Ontologies

One of the motivations of this research was to facilitate leveraging knowledge between multi-domain personnel. The proposed framework composed of two ontologies, which were designed with the purpose of aiding the leveraging of knowledge between two such domains. One of the ontologies aimed at capturing the semantic and syntactic descriptions of a data source. The other ontology aimed at capturing the description of a variety of analysis methods. Knowledge is being leveraged in two ways through the proposed framework; it is evident in the communication between the data source and through the analysis method ontology when reasoning which method matches the data points set for analysis. Here we see knowledge about the data source and its data points being leveraged to the analysis methods. It is also evident in the methods that have been described in the analysis methods ontology. When the methods are selected as potential analysis methods, the user has the ability of viewing further information about that method without needing to have prior knowledge about it. The descriptions of those methods are done through stakeholders that are experts about those methods; in turn their expertise is applied to a set of data and is shared with the user utilizing the system.

In the end we see the sharing of knowledge between the domain experts of the data set in question, with the domain expert(s) who have described the analysis methods in the

analysis method ontology. Through this ability of leveraging knowledge between these two domains, we are able to meet one of the objectives of the framework; to incorporate bi-directional knowledge sharing between domain-specific methods and other methods requiring knowledge from a different domain.

B. The Value of Ontology-based Applications

This research proposed a framework composed of two ontologies in order to demonstrate the advantage of using ontologies in the design. Using multiple ontologies instead of placing all of the descriptions into a sin- gle ontology allows for greater flexibility in the system and re-use of the ontologies. By having an ontology dedicated to data description, we can easily replace this with a domain-specific ontology if necessary. Someone might want to do this if they already have their data source described in a domain specific ontology. As well, this allows for the data source ontology itself to be connected to domain-specific ontologies, which may help pro- vide further semantics to the data points and the data source itself if the user wants to describe it further then what is provided in the ontology. Having the analysis methods described on its own in a separate ontology also provides similar advantages; it provides the system with the flexibility of replacing the entire ontology itself if required. Conversely, the ontology can act directly like a middle layer between the method descriptions and the systems or services that provide access to the methods.

Another advantage of designing an ontology-based application lies with the reasoning ability that the ontology provides to the application. The prototype provided of the proposed framework currently reasons based off of object properties that describe the data points involved and the needs of the methods when soliciting method selection and the ranking of the methods.

The use of ontologies also minimizes the need of having a large infrastructure in an application. In this case, we have the data source and the methods described in the ontology, there is no need to maintain these descriptions within the memory of the application itself, when needed the descriptions can be fetched directly from the ontologies itself.

VII. Acknowledgements

Gabriela Henriques would like to acknowledge Deborah Stacey for her help and guidance throughout this research. Both authors would like to acknowledge the Guelph Ontology Team for their support, as well as the Public Health Agency of Canada for providing the data set used in this research.

References

[1] Buckeridge, D L and Graham, J K and O'Connor, M J and Choy, M K and Tu, S and Musen, M A, *Knowledge-based Bioterrorism Surveillance*, Annual Symposium proceedings, AMIA, 2002.

[2] Buckeridge, D L and Okhmatovskaia, Anna and Tu, Samson and O'Connor, Martin and Nyulas, Csongor and Musen, Mark A, *Understanding Detection Performance in Public Health Surveillance: Modelling Aberrancy-detection Algorithms*, Journal of the American Medical Informatics Association, 2008.

[3] Crubezy, M and O'Connor, M and Pincus, Z and Musen, M A and Buckeridge, D L, *Ontology-centered Syndromic Surveillance for Bioterrorism*, Intelligent Systems, IEEE, 2005.

[4] Firm, T and Chain, V and Network, V, *Ten Ways to Leverage Knowledge for Creating Value*, Knowledge Creation Diffusion Utilization, 2000.

[5] Gómez-Pérez, Asunción and Corcho, Oscar, *Ontology Languages for the Semantic Web*, Intelligent Systems, IEEE, 2002.

[6] Guarino, N and Oberle, D and Staab, S, *What is an Ontology?*, Handbook on ontologies, Springer, 2009.

[7] Henriques, G and Stacey, D A, *An Ontology-based Framework for Syndromic Surveillance Method Selection*, KEOD, 2012.

[8] Keim, D A and Mansmann, F and Schneidewind, J and Ziegler, H, *Challenges in Visual Data Analysis*, Tenth International Conference on Information Visualization, 2006.

[9] Lu, H and Zeng, D and Trujillo, L and Komatsu, K and Chen, H, *Ontology-enhanced Automatic Chief Complaint Classification for Syndromic Surveillance*, Journal of Biomedical Informatics, 2008.

[10] Okhmatovskaia, A and Buckeridge, D L and Shaban-Nejad, A and Sutcliffe, A and Kopec, J A and Fines, P and Wolfson, M C, *SimPHO: An Ontology for Simulation Modelling of Population Health*, Simulation Conference (WSC), IEEE, 2012.

[11] OConnor, M J and Buckeridge, D L and Choy, M and Crubezy, M and Pincus, Z and Musen, M A, *BioSTORM: A System for Automated Surveillance of Diverse Data Sources*, Annual Symposium proceedings, AMIA, 2003.

[12] Provost, F and Hill, S and Bernstein, A, *Towards Intelligent Assistance for a Data Mining Process*, Computer Society, IEEE, 2005.

[13] Sheth, A P, *Changing focus on interoperability in information systems: from system, syntax, structure to semantics*, Interoperating geographic information systems, Springer, 1999