Ontology-based information system for collecting electronic medical records data

Taras Zavaliy, Iouri Nikolski

Abstract – **In this paper questions of data acquisition for intelligent data analysis are considered. The authors describe ontology-based approach for data modeling and management. The ontology sets the domain data structure which can be used in the analysis process. "Imunoskryn" information system for centralized collecting and storing medical data in immunology is briefly described.**

Keywords – **data acquisition, data mining, ontology, OWL, electronic medical records.**

I. INTRODUCTION

There is a need for developing large-scale multivariate models based on the data collected in applied domains of knowledge. The openness and reuse of such data is of great importance, especially in the Web. It is achieved by developing open, semantically aware, adaptive information systems and services. These kinds of systems make it possible to aggregate large amounts of data for statistical or intelligent analysis. Knowledge acquired from this analysis is used in the decision making process. And the quality of decisions greatly depends on the quality of the primary data set.

The authors consider the problem of data acquisition in the overall data mining process. The proposed solution uses domain ontology to set the data structure and aid the data preparation before applying data mining algorithm.

II. DATA ACQUISITION PROBLEM

The task of data acquisition is often undervalued in data analysis. The tasks of acquiring, structuring, integrating and managing potentially erroneous and scarce data from heterogeneous sources are still to be solved. These problems arise on the first three stages of data analysis process. The process itself has five stages [1]: 1) data acquisition; 2) data filtering; 3) data preparation; 4) data mining 5) interpretation. Applied tasks include the design and implementation of data structures and data entry forms, development of tools for error elimination, automated data acquisition, data transformation and integration, building unified data tables before running the mining algorithms.

Subjective bias, human errors, data scarcity, redundancy and inconsistency are often introduced in the data acquisition stage. Relevant attributes of objects are defined by the experts, and the set of these attributes may change when the experts change. Therefore adaptive data structure is needed for collecting large amount of data during long period of time or from different sources.

Let us define main bottlenecks of data acquisition:

1. Data are gathered from different sources;
2. Data are collected using different methodologies;
3. Set of relevant attributes is often changing in time and from source to source.

Suppose every data source uses its own domain ontology. We can thus correlate the bottlenecks of data acquisition to the types and causes of ontology heterogeneity [2] and then use different ontology matching techniques for overcoming it.

III. THE INFORMATION SYSTEM

We use simple domain ontology for defining the structure of attributes in the data table *T*. The ontology is used to eliminate data acquisition problems; facilitate data filtering and preparation. The simplest case of this approach is using the unified vocabularies of terms in order to eliminate data entry errors. In more complex scenario we construct a domain ontology $O = \langle X, R, I \rangle$ that includes hierarchical and semantic relationships between concepts and attributes.

When the set of relevant attributes or their structure is changed a new version of ontology is created. Ontologies from different data sources may also differ. In general the set of ontologies $U_O = (O_1, O_2, \ldots O_i)$ represent modifications of data structures. For integration purposes we define several main operations with ontologies – union, intersection, subtraction and subsumption.

The data structure of the domain can be mapped to the relational form. Each ontology concept is stored in relational table with the unique identifier. Relationships between concepts are stored in the dedicated relational table. This way the relational database can store the data and the adaptive data structure separately.

We used the Protégé editor [3] to create the OWL-ontology for the immunology domain. It defines 4 basic concepts (*Person*, *Hospital*, *Diagnosis* and *Medication*) and a hierarchy of subconcepts. Diagnosis hierarchy is based on the ICD-10 standard. Currently, ontology holds more than 1000 terms. The ontology is a component of "Imunoskryn" web-based information system. The system is used at the immunology departments for patient's data input, medical record tracking and administrative reporting.

IV. CONCLUSION

It was shown that ontology can be used to a) set the domain structure; b) integrate heterogeneous data; c) facilitate data management and analysis. Authors plan further research in these directions and practical applications in medicine.

REFERENCES

[1] S. Mitra, S. K. Pal, P. Mitra, "Data mining in soft computing framework: a survey", IEEE Transactions on Neural Networks, Vol. 13, 2002, pp. 3–14.

[2] Jerome Euzenat, Pavel Shvaiko, Ontology matching. Springer-Verlag, Berlin, 2007, 333 p.

[3] http://protege.stanford.edu. (Cited: 25.11.2009).