

CS594 Midterm Report

Guanrao Chen(gchen@cs.uic.edu)

William Sunna(wsunna1@uic.edu)

Xiaosen Li (xsli2@uic.edu)

Advisor : Dr. Isabel Cruz

03/05/2002

Model-based Mediation with Domain Maps

Guanrao Chen, William Sunna, Xiaosen Li
University of Illinois at Chicago

Abstract

In this project, we propose to use Model-based Mediation with Domain Map to solve the biological problem of integration heterogeneous biological data sources. Compared with other existing method, Model-based Mediation with Domain Map lifts the data output from the syntax level to the conceptual level by using some Conceptual Model Wrappers. Then by using some CM-plugin mechanism, variant CM formalisms can be expressed in one uniform format as a Generic Conceptual Model. The Integrated View Definition uses this GCM and Domain Map which is constructed by domain expert to process the user queries. We use RDF as the GCM and construct the Domain Map with Protégé. Our example queries and results show that this model is an effective approach to deal with those biological problems whose data sources seem disjoint but in fact are related.

1. Introduction

“Complex multiple worlds” refers to the heterogeneous databases which are structurally isolated, but semantically related. Using model based mediation, a query can be answered after correlating the retrieved information from different databases. This database programming technique has particularly important application among the Biology databases.

Huge amount of Biology research data have been generated rapidly last decade. These data are obtained from different disciplines, different species, using different techniques and are stored in different, structurally isolated databases, - “Complex Multiple Worlds”. In order to take full use of these information, model-based mediation is employed to lift the concepts from the database structural level to the conceptual level. A “domain map” which encodes the implicit semantic relationships between the concepts is used as the guidance to navigate through the concepts. This new database programming technique overcomes the isolation of heterogeneous databases and correlates the stored information.

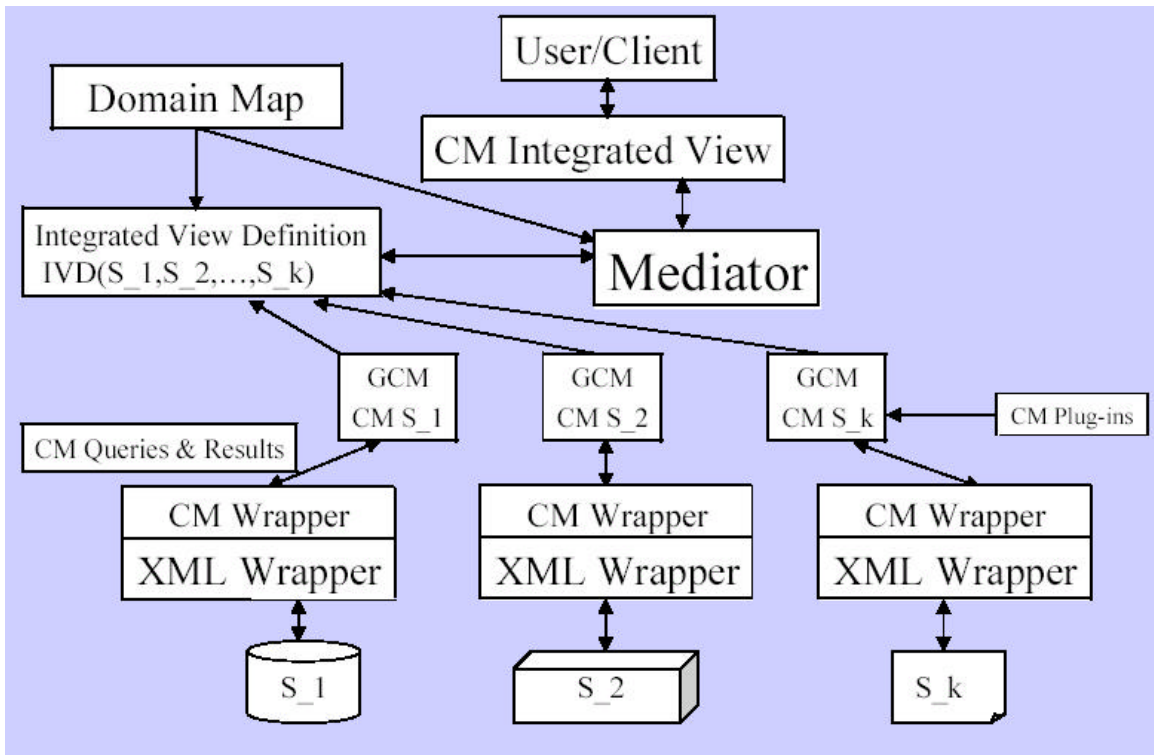
In order to demonstrate this technique, we have chosen two structurally isolated Biology databases, National Center for Microscopy and Imaging Research (NCMIR) and Senselab. The NCMIR database contains 3-dimensional images of neurons, while the Senselab database has

properties about the individual neurons. Our query needs to talk to both databases and correlate their information before the answer can be achieved. Based on our knowledge in Neuroscience, a domain map will be constructed. The concepts will be lifted from the XML database structural level to the RDF conceptual level (Generic Conceptual Model). The domain map will guide us navigate through the knowledge at the conceptual level.

2. Related Work

In recent years, many approaches were proposed to deal with the problem of integrating heterogeneous data sources to elicit information that the individual sources cannot provide independently. B. Ludascher et al proposed an extension to view-based mediation systems called model-based mediation. In this model, views are defined and executed at the conceptual level instead of at the structural level. They also introduced the domain map, semantic net of concepts and relationships that are used to mediate among different data sources, into the model to derive virtual relations. E. Leclearcq et al proposed a semantic mediation approach in the ISIS project to support GIS interoperability. They proposed a loosely coupled architecture based on multi-agent paradigm to share spatial information and services. They also provide a spatial OO model to model distribution and resolve semantic heterogeneities. H . Jamil et al developed an SQL3 compliant database query language called Genomic Query Language for manipulating globally distributed genomic databases. The web-based interface for GQL is capable of understanding the schema of any participating database, assimilating the syntactic and semantic information into the global view and responding to user queries.

3. Conceptual Model



As shown above, in the model-based mediation, data from different sources is first processed by some XML wrappers. Then the CM wrapper lifts the output from syntactic level to conceptual level, which is more semantically meaningful. By using CM plug-ins, variant CM outputs can be converted into one Generic Conceptual Model that makes it is easier for the IVD to deal with. The domain map is constructed by domain experts and is used by the IVD and the mediator to infer domain knowledge.

4. Implementation

- System Overview (Done)
- Elicit Data from NCMIR and Senselab (Done)
- Build data bases in XML for data collected from NCMIR and Senselab (Done)
- Build a domain map (Done)
- Formalize domain map (To be finished)
- Subsystem breakdown and implementation (To be finished)
- Design an algorithm that uses Domain Map to create GCM from XML databases (To be finished)

- Use RDF as the GCM (To be finished)
- USE RQL to query the GCM (To be finished)

5. Conclusion & Future work (To be finished)

Acknowledgment:

We would like to thank Dr. Bertram Ludäscher for giving us information about the Conceptual Model Wrapper , we also would to thank Dr. Isabel Cruz for her guidance through our project.

References

- [1] Model –based Mediation with Domain Maps, B. Ludäscher, A. Gupta, M. E. Martone, *17th Intl. Conference on Data Engineering*, Heidelberg, Germany, IEEE Computer Society, April 2001.
- [2] Model-Based Information Integration in a Neuroscience Mediator System, B. Ludäscher, A. Gupta, M. E. Martone, *demonstration track, 26th Intl. Conference on Very Large Databases (VLDB)*, Cairo, Egypt, September 2000.
- [3] Achieving interoperability of genome databases through intelligent web mediator, H. M. Jamil, *In Proceedings of the IEEE International Symposium on Bio-Informatics and Biomedical Engineering (BIBE 2000)*, Washington, DC, November 8-10, 2000.
- [4] ISIS: A Semantic Mediation Model and an Agent Based Architecture for GIS Interoperability, Eric Leclercq, Djamal Benslimane and Kokou Yétongnon, *In Proceedings of the 1999 International Database Engineering and Applications Symposium, IDEAS 1999*, 2 - 4 August, 1999, Montreal, Canada.
- [5] Associated Biological Information Retrieval From Distributed Databases, Mousheng Xu, Susan Gauch, *The 7th International Conference on Information and Knowledge Management (CIKM '98)*, Washington, D.C., Nov 3-7, 1998.

Work so far

- 1 System Overview
- 2 Elicit Data from NCMIR and Senselab
- 3 Build databases in XML for data collected from NCMIR and Senselab
- 4 Build a domain map

Work remains to be done

- 1 Formalize domain map
- 2 Subsystem breakdown and implementation
- 3 Design an algorithm that uses Domain Map to create GCM from XML databases
- 4 Use RDF as the GCM
- 5 USE RQL to query the GCM

Questions on how to proceed

- 1 Use Protégé to built the Domain Map or write a C++ program to do the job

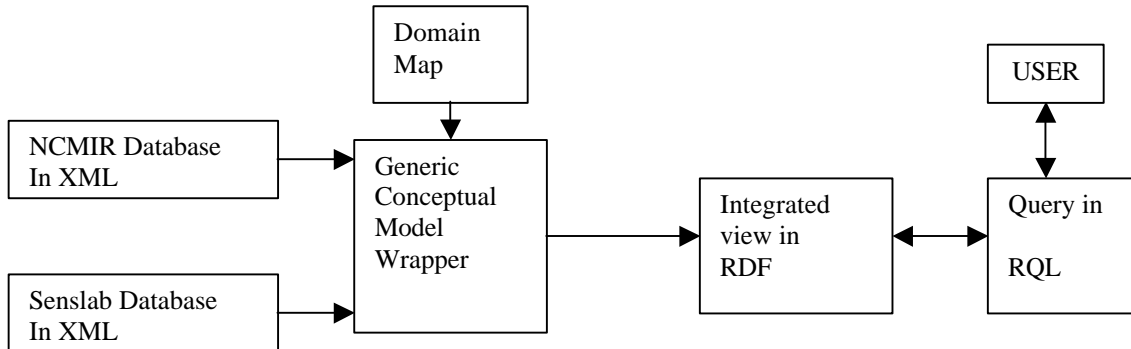
We found out (After trying to implement the Domain map in Protégé) that Protégé is good for describing classes, Properties, Subclasses, Inheritance, Constraints, define instances, overriding....etc. Our domain map goes beyond that. We need to define relations between the various classes and how they interact between each other. Our Domain map is a Digraph and we are wondering if we missed something in Protégé that can help us realize the Digraph structure of our domain map

- 2 Given time constraint we don't think we will be able to implement the Generic Conceptual Model Wrapper subsystem (Done by hand in the original reference paper – Please refer to Appendix A). Instead we are thinking of describing this by using pseudo code on how this subsystem work. Can we proceed with that?

- 3 We collected some good data to work with (low scale example). This data is sufficient in our opinion to demonstrate how the mediation system works. We are hoping to generalize our findings to be used on large-scale databases. Can we proceed with our small database (found on the web site: <http://www.cs.uic.edu/~wsunna/594/594.html>) or enlarge the databases.

Appendix A: Example Application

To illustrate the process of model based mediation with domain maps we will navigate through an example. We will work on two databases taken from NCMIR (Neuron image database) and Senselab database. Here are the main blocks of our system:



The two databases in the XML look like this:

1. Senselab (Part of the data base in XML):

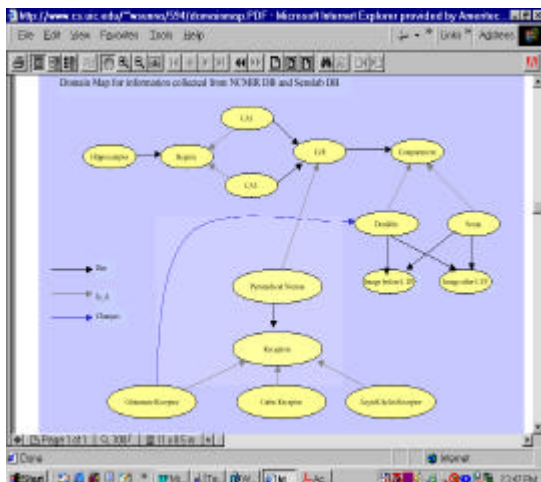
```
<?xml version="1.0" ?>
<Hippocampus>
  <Regions>
    <Region id="CA1">
      <Cell type = "Pyramidal
Neuron" >
        <Receptors>
          <Glutamate_Receptor>
Y </Glutamate_Receptor>
          <Gaba_Receptor> Y
        </Gaba_Receptor>
      </Cell>
    </Region>
  </Regions>
</Hippocampus>
```

2. NCMIR (Part of the data base in XML):

```
<?xml version="1.0" ?>
<Hippocampus>
  <Regions>
    <Region id="CA1">
      <Cell type = "Pyramidal
Neuron" >
        <Receptors>
          <Glutamate_Receptor>
Y </Glutamate_Receptor>
          <Gaba_Receptor> Y
        </Gaba_Receptor>
      </Cell>
    </Region>
  </Regions>
</Hippocampus>
```

For complete XML files, please visit: <http://www.cs.uic.edu/~wsunna/594/594.html>

The domain map looks like this:



The domain map will be implemented using Protégé. The Generic conceptual Wrapper will take the two XML files as an input and use the domain map to build and RDF schema at GCM (Generic Conceptual Model) that can be queried by RQL.

We chose this query to demonstrate the need for both data bases to be integrated and the need for a domain map to be used:

Does hippocampus CA1 pyramidal cell has glutamate receptor and what the effect on the neuron shape caused by the Ca++ ion entered through the glutamate receptor during LTP?