Preferences for Queries in a Mediator Approach

Alain Bidault and Sarah Cohen Boulakia and Christine Froidevaux[1]

Abstract. The problem of integrating relevant information obtained from multiple heterogeneous sources is a complex task, with which biologists are now faced. In this paper, we address the problem of querying biomedical databases in a mediator context. We propose to exploit the metadata of the sources to take into account user preferences. The mediator system we present is designed within a tractable logical framework. It allows both transparent and cooperative querying and makes it possible to keep track of the origins of the instances provided as answers. Our proposal is generic in that it is relevant not only for bioinformatics, but could also be applied to other domains for which metadata are available.

1 Introduction

With the increasing amount of disparate biomedical data, there is now a clear need for interoperability between sources in bioinformatics. Several attempts to integrate biomedical data have been made in the academic and industrial sectors: portals (e.g. SRS [12], Entrez [1]), platforms (e.g. ISYS [11], Genostar [1]) data warehouses and mediators [9] (e.g. TAMBIS [4], Biomediator [10], GIMS, GUS [1]).

In this work, we investigated ways of taking into account user preferences in a mediator context. This context makes it possible for users to focus on specifying their demand, releasing them from the necessity of having to find the relevant sources and possibly of combining data from multiple sources to obtain answers. It is also well adapted to the need for frequent updates of biomedical databases. Many information integration systems have been developed in the last ten years (e.g. PICSEL [7], Information Manifold, SIMS, TSIMMIS [1]).

Because we need (i) an industrially relevant integration system with (ii) a strong theoretical basis and (iii) a language readily understood by users and expressive both for queries and for source descriptions, we chose to work in the context of the PICSEL project. A complete version of the software will be released by France Telecom in 2004. Furthermore, PICSEL has a *Local As View* approach convenient for easy updates of the sources.

We aimed to develop an extended mediator system allowing both *transparent* queries (as usual) and a **cooperative** answering process, crucial in the biomedical area but missing in the integration systems developed within the bioinformatics community to date.

2 The Domain Knowledge and the Sources

The domain knowledge is expressed in the spirit of that in PICSEL by means of a declarative representation of classes (Chromosome, Sequence, ...) and of relationships between classes. It is described using atoms of the form $p(\bar{X})$, where p is a relation name and $\bar{X}$ a tuple of variables. In the set $\mathcal{E}_c$, we distinguish some unary relations, called *concepts*. This set contains the set $\mathcal{E}_{cd}$, the classical set of concepts of PICSEL. We call $C(x)$ an *atom-concept* if C is a concept. We use standard first-order logic semantics.

The domain knowledge $(\mathcal{D}, \mathcal{C})$ contains two components:

- A set $\mathcal{D}$ of **rules** of the form: $p_1(\bar{X}_1) \wedge \ldots \wedge p_n(\bar{X}_n) \Rightarrow q(\bar{Y})$, where $p_i(\bar{X}_i)$ and $q(\bar{Y})$ are atoms, with $\bar{Y} \subseteq \cup_{i\in[1..n]}\bar{X}_i$.

The rules in $\mathcal{D}_h \subseteq \mathcal{D}$ describe a *hierarchy* of the domain concepts and are of the form: $C_1(x) \Rightarrow C_2(x)$, where C_1 and C_2 are concepts of $\mathcal{E}_c$. Concepts and relationships not appearing in rules as conclusions are called *base relations*.

- A set $\mathcal{C}$ of **constraints**, $c : l_1(\bar{X}_1) \wedge \ldots \wedge l_n(\bar{X}_n) \Rightarrow \bot$ where $l_1(\bar{X}_1), \ldots, l_n(\bar{X}_n)$ are literals, with at most one negative.

An expansion step is a backward chaining step on the rules of $\mathcal{D} \cup \mathcal{C}$ and generates a **rewriting** of the user's query $Q(\bar{X})$.

Contents of a source $\mathcal{S}_i$ are represented, as in PICSEL, by means of a vocabulary $\mathcal{V}$ containing as many local relations v_{ij}, called *views*, as we know the source $\mathcal{S}_i$ gives instances of j domain relations. The description of sources in terms of views has two components:

A logical set of implications $\mathcal{D}_v \subseteq \mathcal{D}$ that link each view to a domain relation, $v_{ij}(\bar{X}) \Rightarrow p(\bar{X})$. **ex:** $v_{11}(x, y) \Rightarrow IsLocatedIn(x, y)$ $v_{21}(x) \Rightarrow Protein(x)$ (Full example can be found at [2]).

A set of constraints $\mathcal{C}_v \subseteq \mathcal{C}$ characterizing the view instances: $l_1(\bar{X}_1) \wedge \ldots \wedge l_n(\bar{X}_n) \Rightarrow \bot$, where $l_1 \ldots l_n$ are base relations and/or view names or their negation. **ex:** $v_{11}(x, y) \wedge \neg InChromo(y) \Rightarrow \bot$

3 Metadata for User Preferences

Metadata is data that characterizes other data in a reflexive way, e.g., data to describe the quality of the contents of the sources. To facilitate the definition of metadata about the **user's preferences**, we introduce a new set of rules $\mathcal{D}_P = \mathcal{D}_s \cup \mathcal{D}_l \cup \mathcal{D}_q \cup \mathcal{D}_f$.

In $\mathcal{E}_c$, we introduce three classes of concepts to express properties on the sources: $\mathcal{E}_{cs}$ contains the predicates that are names of sources, $\mathcal{E}_{cq}$ is made of concepts that define the reliability of the sources and $\mathcal{E}_{cf}$ is dedicated to concepts that define the focus of a source.

$\mathcal{D}_s$ is the set of rules of the form $v(\bar{X}) \Rightarrow Source(x_j)$, with $Source \in \mathcal{E}_{cs}$ *and one rule for each* $x_j \in \bar{X}$, meaning that x_j comes from $Source$, which is the only **source** that contains v.

ex: Let us consider that two sources are available from the mediator: GB (GenBank) and RS (RefSeq). We obtain formulas of this form:
$\{v_{11}(x, y) \Rightarrow GB(x),\ v_{11}(x, y) \Rightarrow GB(y),\ v_{21}(x) \Rightarrow RS(x)\} \subset \mathcal{D}_s$.

Although the existing biomedical data banks were designed by different research teams in different contexts and are therefore highly heterogeneous, they are nonetheless related. In particular, biomedical data banks more and more frequently refer to each other by means of hypertext links called **cross-references**. These links may be very

[1] L.R.I., C.N.R.S U.M.R. 8623, & University of Paris-Sud Bâtiment 490, 91405, Orsay Cedex, France {bidault, cohen, chris}@lri.fr

useful in that they make it possible to obtain additional information concerning a single instance of one entity in a given source by providing access to more detailed information in other sources. It is worth noticing that these links are not symmetrical in biomedical websites. $\mathcal{D}_l$ is the set of rules that express **links** from one database to another. It has two subsets: rules in $\mathcal{D}_k$ specify whether a given source **knows** another source and rules in $\mathcal{D}_{cr}$ handle **cross-references**.
ex: $RS(x) \wedge GB(y) \Rightarrow Knows(x, y)$ *RefSeq knows GenBank.*
$v_{21}(x) \wedge v_{11}(y, z) \Rightarrow CrossRef(x, z)$ *There is a cross-reference from Protein of* v_{21} *in RS to the second argument of* v_{11} *in GB.*
Beside, we have introduced a symmetric predicate, "TheyKnow", which is potentially more intuitive for biologists. The two predicates are related as follows:
$Knows(x, y) \wedge Knows(y, x) \Rightarrow TheyKnow(x, y)$
$TheyKnow(x, y) \Rightarrow Knows(x, y) \wedge Knows(y, x)$.

In the biomedical domain [1], two kinds of sources are usually distinguished: *primary databases*, which provide raw data such as sequences submitted by various laboratories (GenBank) and *secondary databases*, which contain only data that have been validated (Swiss-Prot, LocusLink). Moreover, as biomedical data often reflect the personal views of experts, each biologist implicitly assigns to the data in the sources a level of reliability, depending on the confidence she/he has in the sources. Thus, the reliability of the instances returned as answers for a query depends on the database consulted.
$\mathcal{D}_q$ is the set of rules of the form $v_i(\bar{X}) \Rightarrow ConceptQ(x_i)$, with $at\ most\ one\ rule\ for\ some\ x_i \in \bar{X}$, which provides information about the **reliability** of the instances of the argument x_i of the view v_i through the concept $ConceptQ \in \mathcal{E}_{cq}$.
ex: $v_{11}(x) \Rightarrow ALPoor(x)$ *instances of* v_{11} *are* at least *of poor reliability.*.

The owners of biomedical data sources are generally experts in a specific biomedical domain and offer their own point of view on their source by focusing on a specific biomedical entity. The **focus** of a source is defined as the entity around which the source is organized. Querying a database focusing on a single entity ensures that the biologist can obtain precise information concerning that entity.
It should be stressed that the declaration of a focus relates to a source and not to a view. Moreover, if a variable is used in arguments of different predicates from different sources, it may be associated with different focuses, with no more than one focus for each source.
$\mathcal{D}_f$ is the set of rules of the form $Source(x) \Rightarrow FocusConcept(x)$ where $FocusConcept \in \mathcal{E}_{cf}$ and $FocusConcept$ is defined as the **focus** of $Source$, with at most one focus per source.
ex: $GB(x) \Rightarrow FocusNuclSeq(x)$ *focus of GB is the nucleotide Sequence.*

It is often important for biologists to know the origin of the answers given by the mediator because they do not have the same confidence in all sources. The instances returned by the rewritings $Q_{\mathcal{R}j}(\bar{X}), j \in [1..k]$, with k the **number of the terminal rewritings**[2] of $Q(\bar{X})$, are therefore grouped and presented as a set $\mathcal{E}_{\mathcal{P}} = \{\mathcal{E}cpl_1, \ldots, \mathcal{E}cpl_k\}$, where $\mathcal{E}cpl_j$ is a set of couples (set of sources s_{ji}, variable x_i) with $i \in [1..n_x]$. This indicates that instances of x_i come from all the sources of s_{ji} for some given rewriting[3].

Definition 3.1: *Let* $Q(\bar{X})$ *be a query and* $Q_{\mathcal{R}j}, j \in [1..k]$, *one of its terminal rewritings. Let* $\mathcal{V}_j = v_{j1}, ..., v_{jnj}$ *be the set of the* n_j *views*[4] *that appear in* $Q_{\mathcal{R}j}$. *Consider* $x_i \in (\bar{X})$, $i \in [1..n_x]$, *and* $\mathcal{V}_{jxi} \subseteq \mathcal{V}_j$ *the set* $\bigcup_{l \in [1..n_{ji}]} v_{ijl}(xi, \bar{y})$ *of the views where* x_i *appears. Let* $\mathcal{E}_{Sjxi}$ *be the set of all the source names* $\mathcal{S}_{ji}$ *such that there exists a rule* $v_{ijl}(xi, \bar{y}) \Rightarrow \mathcal{S}_{ji}(xi)$. *We call* $\mathcal{E}_{\mathcal{P}}$ **the presentation set** *of* $Q(\bar{X})$, $\mathcal{E}_{\mathcal{P}} = \bigcup_{j \in [1..k]} \bigcup_{i \in [1..n_x]} (x_i, \mathcal{E}_{Sjxi})$.

Starting from these results, we are extending our previous algorithms [5] so that they calculate in a tractable way the terminal rewritings of a query containing metadata. Later, this algorithm will integrate some new features (e.g. accessibility metadata, synonyms).

4 Comparison with Other Approaches

Many fruitful discussions with biologists have emphasized the need to improve classical integration systems by introducing metadata. It is clear that the biomedical application domain should benefit from a mediator approach which would add features that fit the specific nature of biomedical data (e.g. reliability of the data depending on the sources). [6] has proposed an algorithm that permits to select automatically sources to be queried according to user's preferences. Here, we have presented a formalism that allows the user to access the sources offered by a mediator system in a transparent manner and to obtain information about the sources accessed. Our proposal is generic as far as it could also be applied to other domains.
Our framework is original in that metadata are expressed in the same logical language as the data, but are used in a specific way. This is different from the DublinCore [3] and SEMEDA [8] approaches that do not offer the possibility of using metadata in an expressive and simple query language. Moreover they do not provide a way of defining quality criteria specific to biomedical data at all.

Acknowledgement We would like to thank S. Lair and N. Stransky (Curie Institute), B. Labedan (IGM, Orsay) and the partners of the AS-STIC CNRS for fruitful discussions. This work was supported in part by a French Research Action CNRS AS-STIC [1].

REFERENCES

[1] All references at http://www.lri.fr/~bidault/articles/ref_ecai04.html
[2] Full example at http://www.lri.fr/~bidault/articles/example_ecai04.html
[3] http://dublincore.org/documents/1998/09/dces
[4] P.G. Backer, C. Goble, S. Bechhofer, N.W. Paton, R. Stevens, A. Brass. An ontology for bioinformatics applications. In *Bioinformatics*, 15(6), pp. 510-520, 1999.
[5] A. Bidault, C. Froidevaux, B. Safar. Similarity Between Queries in a Mediator Approach. In *Proc. of ECAI*, pp. 235-239 2002.
[6] S. Cohen Boulakia, S. Lair, N. Stransky, S. Graziani, F. Radvanyi, E. Barillot, C. Froidevaux. Selecting Biomedical Data Sources according to User Preferences. In *Proc. of ISMB/ECCB*, 2004.
[7] F. Goasdoué, V. Lattès, M.-C. Rousset. The Use of CARIN Language and Algorithms for Information Integration: The PICSEL Project. In *Int. J. of Cooperative Inf. Syst.* (IJCIS), 1999.
[8] J. Kohler, S. Philippi, M. Lange. SEMEDA: ontology based semantic integration of biological databases. In *Bioinformatics*, 19(18), 2002.
[9] A. Y. Levy. Combining Artificial Intelligent and Databases for Data Integration. In *Artificial Intelligence Today*, pp. 249-268, 1999.
[10] R. Shaker, P. Mork, M. Barclay, P. Tarczy-Hornoch. Rule Driven Bi-Directional Translation System for Remapping Queries and Result Sets Between a Mediated Schema and Heterogeneous Data Sources. In *Proc. of AMIA An. Symp*, 2002.
[11] A.C. Siepel, A.D. Farmer, A.N. Tolopko, M. Zhuang, P. Mendes, W. Beavis, B. Sobral. ISYS: a decentralized, component-based approach to the integration of heterogeneous bioinformatics ressources. In *Bioinformatics*, 17(1), pp. 83-94, 2001.
[12] E. M. Zdobnov, R. Lopez, R. Apweiler, T. Etzold. The EBI SRS server - recent developments. In *Bioinformatics*, 18(2), pp. 368-373, 2002.

[2] a rewriting is terminal if it is a conjunction of atom-views
[3] n_x is the size of the distinguished variables vector $\bar{X}$
[4] For a query Q, the number of predicates of its rewritings may differ from one rewriting to another.