Personalization through Specification Refinement and Composition

Dmitry O. Briukhov, Leonid A. Kalinichenko, Nikolay A. Skvortsov
Institute for Problems of Informatics, Russian Academy of Sciences
E-mail: {brd,leonidk,scvora}@synth.ipi.ac.ru

Abstract

Issues of personalized digital collections design using heterogeneous information sources registered at the subject mediating environment are considered. The approach provides for design of collections satisfying personal needs of end users in terms of information content and representation. A personalized collection is formed as a composition of relevant fragments conformant to the mediated schema. To do this, a method for the compositional design of information systems [2] is applied.

1 Introduction

Mediation of heterogeneous information sources provides an approach for intelligent information integration [16]. Important application areas greatly benefit from the *subject mediation* approach supporting information integration in a particular subject domain. Among them are Web information integration systems, digital libraries providing content interoperability, digital repositories of knowledge in certain domains (like: Digital Earth, Digital Sky, Digital Bio, Digital Law, Digital Art, Digital Music). This technology is considered as a promising alternative to the widely used general purpose Web search engines characterized by very low precision of search due to uncontrollable use of terms for indexing and retrieval.

The approach developed in this paper is aimed at design of virtual collections in a mediating environment satisfying specific information requirements of users. The requirements are supposed to be formulated in the form of specifications of a domain defining the user's conceptual and terminological context and of collection specifications that should be created to satisfy the needs of a particular user or a group of users. To improve the semantic contents of the personalized collection specification, its elements are associated with concepts in the domain specifications.

In this paper an approach, known as Local as View (LAV), is assumed for the mediator organization. LAV considers schemas exported by heterogeneous information sources as materialized views over virtual classes of the mediated schema defining mediator itself. The mediated schema includes ontological/terminological definitions of mediator's subject domain. Queries are expressed in terms of the mediated schema. Query evaluation is done by query planning making its rewriting in terms of the source schemas. It is assumed here that all specifications (including mediated schemas) are expressed in terms of a canonical information model SYNTHESIS [10].

Works on information personalization have been published in different sources [12, 14, 5, 7, 4]. Personalization design methods proposed in this paper are based on the principles of information reuse and compositional design of information systems. Fragments of source collection specifications refining [9] fragments of mediated schema are composed to form specifications of data accessible through the mediator. Analogously personalized requirement specifications must be refined by fragments of specifications of the mediated schema. Ontological/terminological context definitions of a mediator should be contextualized in the personalized context. After that relevant elements of mediated schema specifications are found in order to discover specification fragments refining specifications of requirements. Then, the personalized collection specification is defined as a composition of the fragments found.

Note that mediation environment is recursively constructed: a mediator can be registered at another mediator as a collection. In particular, this may lead to composition of several subject domains in a mediator. Therefore applying here our approach to a mediated schema does not limit the generality and scale of the method.

Implementation of the design approach proposed is based on the SYNTHESIS CASE-system prototype developed for compositional design of information systems [2].

2 Organization of Metainformation in the Mediating Environment

The mediator is organized as a set of facilities over heterogeneous information sources (collection registration, browsing, querying, data extractions, etc.). Any specification in the mediator is represented in the canonical information model and is stored in the metainformation repository. The metainformation repository includes the following levels of specifications:

- *local* level for specification of information sources registered at the mediator together with their contexts;
- *mediated* level for definition of the mediator's subject domain;
- *personalized* level for specifications of user's information requirements.

Information sources registered at the mediator are heterogeneous - such as databases, unstructured textual collections, semistructured web sites etc. When a collection is being registered at the mediator it is important to extract related metainformation characterizing unstructured as well as structured data. Underlying semantics of the registered information can be defined applying ontological modeling. The structure of semistructured information should also be revealed during registration. For instance, the structure of HTML-sites can be revealed by analyzing tags of hypertext documents and recognizing regularities in their structure, XML data structure is described using the data type definitions (mapping of the XML standard data models into the canonical model is considered in [11]).

The SYNTHESIS model [8] supported by the mediator's metainformation repository is used to define a subject domain and to construct a uniform representation of heterogeneous information sources and manipulate them in a uniform manner. This model provides for homogeneous description of heterogeneous information sources, structured (databases), semistructured (hypertext documents), and unstructured (text documents) data, as well as for conceptual definitions of subject domains (applying ontological specifications, thesauri, vocabularies, classifiers), activities and workflows. In SYNTHESIS, frames are used as autonomous self-defined entities representing unstructured and semistructured data. The typed (object) model of the language is built above frames. Classes in SYNTHESIS represent sets of object instances from the subject domain.

The specification calculus [9] is used to support operations over the mediator's metainformation. These operations are intended for specification manipulations:

- *reduct* operation defines a subset of type specification as its supertype.
- *meet* operation yields the intersection of two types as the most common reduct of the operand types specifications.
- *join* operation yields the union of specifications of the operand types.

Compositions of classes are based on operations over collections of instances. Such compositions may form views over classes.

3 Personalized Information Requirements Specification

A procedure is assumed for registering users or groups of users to define their information requirements. At a mediator users may define their requirements by a combination of the following:

- expanding the subject domain thesaurus and a list of classifier categories for relating information available to them;
- specifying terms related to their area of interest and select the list of classifier categories to which their data of interest belong;

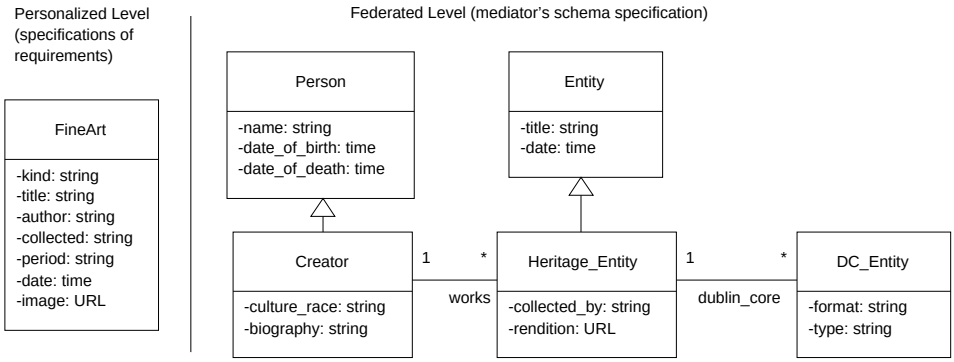

Figure 1: Example

- defining structure of required information;
- specifying own information representation and inference;
- including specific services additional to services defined for a subject mediator.

Expanding or selecting of subset of mediated ontological/terminological concepts or classifier categories leads to a required extension (narrowing) of content of user queries. Information structuring is expressed in a form of collection schema designed as a view over the mediated level. Thus, a user can specify the content and the form of information of interested. Personalization specification includes definitions of ontological/terminological concepts for users to be involved in personalized collection design process.

4 Compositional Design

Personalized collection is designed as a composition of relevant fragments of the mediated schema refining specification of personalized requirements. According to the approach proposed, the development of a collection specification includes the following stages [2].

- Define the required collection specifications in terms of the canonical model. Place specification elements to be instances of concepts of an ontological or terminological contexts of the user model for expressing the semantics of elements;
- Map the mediated ontology into the ontological context of the requirement specifications finding intercontext links;
- Through the intercontext concept links, establish correlation of schema specification elements of the requirement specifications with the respective elements of mediated specifications. Correlated elements are ontologically relevant.
- Among the ontologically relevant specification elements of mediated schema find those that can be used to refine fragments of the requirement specifications. Resolve structural conflicts and mismatches between schemata.
- Design views over classes of mediated schema to implement classes of specifications of requirements.

Figure 1 contains an example of schema specifications at personalized and mediated levels of the repository. The mediated specifications represent a part of the Cultural Heritage mediator schema. Personalized specifications define requirements for the users that need only general information on fine art. The following sections show how to implement specifications of requirements over the mediated schema.

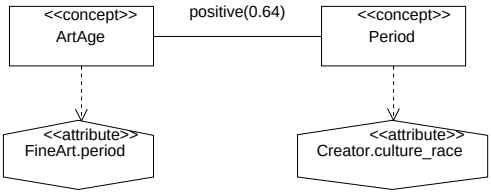

Figure 2: Identification of ontologically relevant elements

5 Mapping of Conceptual Metainformation

All schema specifications are associated with ontological contexts defining concepts of the respective subject domains (in form of thesaurus and/or ontology). Concepts in the canonical model are defined as abstract data types. On the other hand concepts may have properties of thesaurus lexical units. They can be associated by one of several kinds of semantic relationships [1]. Concepts are characterized by their verbal definitions and descriptor lists containing meaningful words from the verbal definitions. Concept descriptors are used for establishing relationships with other concepts outside the given ontological context.

The mediated ontology concepts are mapped into ontological contexts of the personalized collections. At the first stage, relationships between concepts of different contexts are established by calculating correlations between concepts using their verbal definitions. For this purpose, positive, hypernym (hyponym) relationships between concepts of two ontological contexts are calculated using weighted vector-based information retrieval approach [3, 13, 15]. After that a deeper and more accurate concept integration may be performed (if required) using the concept specifications in form of abstract data types. In this case compositional methods are used for concepts reconciliation.

After the mediated conceptual context has been mapped into the personalized one, we pass to the identification of mediated schema elements relevant to personalized specifications (including types, classes, and their fragments). While concept relationship paths are analyzed, the concept graph might be complemented with the missing relationships. Establishing complementary relationships is performed applying the algorithm based on the transitivity property of semantic relations [6]. Relationships inferred are used to identify correspondences between the specification elements. Ontologically relevant may become elements of the specifications of the same kind (type, class, function, attribute, etc.) that are instances of interrelated ontological classes.

For example, let the attribute *FineArt.period* be an instance of the ontological class of personalized concept *ArtAge*, and the attribute *Creator.culture_race* be an instance of the ontological class of mediated concept *ArtAge* (see Figure 2). If the concept *ArtAge* is detected to be positively related to the mediated concept *Period*, then we can assume that the attribute *FineArt.period* is ontologically relevant to the attribute *Creator.culture_race*.

6 Composition of Personalized Collection

The method used to identify fragments of mediated schema refining fragments of specifications of requirements is based on the principles defined in [9].

Fragments of the ontologically relevant types of mediated schema are selected to refine the respective fragments of the collection being developed. The fragments of type specifications are defined applying the type reduct operation. Reducts are considered as patterns of refining specifications. In the process of refining, various conflicts and mismatches in the specification fragments of the mediator and personalized definitions are resolved. Resolving structural conflicts between the mediated schema and personalization requirements is performed in the process of comparing of ontologically relevant paths. Minimal paths must be chosen, i.e. those that do not contain relevant subpaths.

In our example the type *FineArt* is ontologically relevant to mediated type *Heritage_Entity*. Attributes of *FineArt* are relevant to attributes of different types: *FineArt.title* to *Entity.title*, *FineArt.author* to *Person.name* and so on. In case of *FineArt.title* there is no conflict, and it can be immediately refined by *Entity.title*. In case of *FineArt.kind* the path *Heritage_Entity—dublic_core—DCE_Entity—type* is used to refine it. For attributes *FineArt.image* and *Heritage_Entity.rendition* semantic conflict exists, because renditions can be not only images. Thus the function is required for them to choose images from URLs

using the information of Dublin Core specifications (*DC_Entity*) about format of described object.

A pair of reducts is constructed for types *Heritage_Entity* and *FineArt*: common reduct *R_FineArt_Cultural_Heritage* of *FineArt* w.r.t. *Heritage_Entity*, and concretizing reduct *CR_FineArt_Cultural_Heritage* of type *Heritage_Entity* which includes specification how to model *R_FineArt_Cultural_Heritage* reduct attributes though the values of *Heritage_Entity* type attributes. Functions *get_kind* and *get_image* are used to resolve conflicts between types.

Here is an example of concretizing reduct for these types:

```
{ CR_FineArt_Heritage_Entity;
  in: c_reduct;
  metaslot
    of: Canvas;
    taking: {dublin_core, title, collected_by, date, rendition};
    reduct: R_FineArt_Heritage_Entity
  end;
  simulating: {
    R_FineArt_Heritage_Entity.kind      ~ get_kind
    R_FineArt_Heritage_Entity.title     ~ CR_FineArt_Heritage_Entity.title
    R_FineArt_Heritage_Entity.collected ~ CR_FineArt_Heritage_Entity.collected_by
    R_FineArt_Heritage_Entity.date      ~ CR_FineArt_Heritage_Entity.date
    R_FineArt_Heritage_Entity.image     ~ get_image };

  get_kind: {in: function;
    params: {+ext/CR_FineArt_Heritage_Entity, returns/string};
    predicative: { returns = ext.dublin_core.type} };

  get_image: { in: function;
    params: {+ext/CR_FineArt_Heritage_Entity, returns/URL};
    predicative: {
      ( ext.dublin_core.format = 'jpeg' | ... ) &
      returns = ext.rendition} }
}
```

Aanalogously for the pair of types *Creator* and *FineArt* common reduct and and concretizing reduct are constructed, where attributes *FineArt.author* and *FineArt.period* are modeled by *Person.name* and *Creator.culture_race* respectively and attrubute *works* is included for relating to *Heritage_Entity*.

Note that the techniques that have been used so far can be applied as the basis for recommender systems to make personalized recommendations of services in terms of refining specification fragments of the mediated schema and respective classes.

After concretizing reducts of types have been identified views *v_heritage_entity* and *v_creator* are constructed over the mediated classes such that reducts are instance types of these views. Among instances of *heritage_entity* those ones are chosen whose Dublin Core specification says that it is fine art, i.e. *Heritage_Entity.dublin_core.type* equals 'painting', or 'drawing', or 'watercolor'.

To implement personalized class *fineArt* the view *v_fineArt* is constructed as a composition of views over mediated classes. In our case veiws *v_heritage_entity* and *v_creator* must be joined by the attribute *Creator.works*. The view becomes subclass of the class *fineArt* in the requirement specifications.

For correct transformation of types during view manipulations, the *meet* and *join* operations over concretizing reducts are applied to build compositions of reducts. These compositions are intended to provide right instance types for all developed views. In particular, join of views implies applying *join* operation on instance types, union of views implies *meet* of types. More details on the compositional methods can be found in [2, 9]. In the example, concretizing reducts *CR_Heritage_Entity* and *CR_Creator* are composed into the concretizing type *CT_FineArt* using *join*. The resulting type refines *FineArt*.

$$CT\_FineArt = CR\_FineArt\_Heritage\_Entity \sqcup CR\_FineArt\_Creator$$

Thus resulting view *v_fineArt* looks as following:

```
{ v_fineArt; in: class;
    metaslot form_view: { in: function; enforcement: on_access;
      params: -returns/v_fineArt as_class;
      { v_heritage_entity(h) & v_creator(c) & in(h,c.works) } }
    end
  superclass: fineArt;
  instance_section: CT_FineArt
}
```

Conclusion

An approach for design of personalized digital collections at the subject mediation environment using the specifications of user's information requirements has been presented. For this task we apply methods of compositional information systems design. A personalization is defined as a composition of mediated schema fragments refining specifications of user's information requirements.

References

[1] *ISO 5964. Documentation - Guidelines for the Development and Establishment of Multilingual Thesauri.* ISO, 1985

[2] D. O. Briukhov, L. A. Kalinichenko. *Component-Based Information Systems Development Tool Supporting the SYNTHESIS Design Method.* Proc. of the East European Symposium on Advances in Databases and Information Systems, Poland, Lect. Notes Comput. Sci., 1998, no. 1475

[3] D. O. Briukhov, S. S. Shumilov. *Ontology Specification and Integration Facilities in a Semantic Interoperation Framework.* Proc of the International Workshop on Advances in Databases and Information Systems (ADBIS'95), 1995

[4] J. Budzik, K. Hammond, L. Birnbaum. *Information access in context.* Knowledge based systems, Elsevier, 2001

[5] J. Cruz, M. Klink, T. Krichel. *Personal Data in a Large Digital Library.* Research and Advanced Technology for Digital Libraries, 4th Conf., ECDL 2000, Lisbon, Portugal, LNCS 1923, Sep 2000

[6] P. Fankhauser, E. J. Neuhold. *Knowledge Based Integration of Heterogeneous Databases.* Integrated Publication and Information Systems Institute (GMD-IPSI), Darmstadt, 1993

[7] G. Jones, D. Quested, K. Thomson. *Personalised Delivery of News Articles from Multiple Sources.* Research and Advanced Technology for Digital Libraries, 4th European Conference, ECDL 2000, Lisbon, Portugal, LNCS 1923, p. 127, Sep 2000

[8] L. A. Kalinichenko. *SYNTHESIS: the Language for Specification, Design and Programming of the Interoperable Environments of Heterogeneous Information Resources.* Institute for Problems of Informatics, Russian Academy of Sciences, Moscow, 1993

[9] L. A. Kalinichenko. *Compositional Specification Calculus for Information Systems Development.* In Proc. of the East-West Symposium on Advances in Databases and Information Systems (ADBIS'99), Maribor, Slovenia, Lecture Notes Comput. Sci., 1999

[10] L. A. Kalinichenko. *Integration of Heterogeneous Semistructured Data Models in the Canonical One.* Proc. 1st Russian Sci. Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collections", St. Petersburg, 1999

[11] O. Machulsky, M. Osipov, L. A. Kalinichenko. *Mapping the XML Data Model into the Object Model of the SYNTHESIS Language.* Proc. 1st Russian Sci. Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collections", St. Petersburg, 1999

[12] P. Resnick and H. Varian, Guest Editors. *Recommender Systems.* Communications of the ACM, Vol. 40, No. 2, March, 1997

[13] G. Salton, C. Buckley. *Term-Weighting Approaches in Automatic Text Retrieval.* Readings in Information Retrieval, K. S. Jones and P. Willett, Kaufmann, 1997

[14] J. W. Schmidt, G. Schroder, C. Niederee, F. Matthes. *Linguistic and Architectural Requirements for Personalized Digital Libraries.* International Journal on Digital Libraries, Vol 1, No 1, April 1997, Springer, 1997

[15] N. A. Skvortsov, L. A. Kalinichenko. *An Approach to Ontological Modeling and Establishing Intercontext Correlation in the Semistructured Environment.* Proc. 2nd Russian Sci. Conf. "Digital Libraries: Advanced Methods and Technologies, Digital Collections", Protvino, 2000

[16] G. Wiederhold. *Mediators in the Architecture of Future Information Systems.* IEEE Computer, 1992