Mediators over taxonomy-based information sources

Yannis Tzitzikas1,⋆, Nicolas Spyropoulos2, Panos Constantopoulos1

1 Department of Computer Science, University of Crete, Greece and Institute of Computer Science, ICS-FORTH; e-mail: {tzitzik|panos}@csi.forth.gr
2 Laboratoire de Recherche en Informatique, Université de Paris-Sud, France; e-mail: spyratos@lri.fr

Edited by M. Missikoff. Received: November 29, 2001 / Accepted: October 30, 2003
Published online: February 6, 2004 – © Springer-Verlag 2004

Abstract. We propose a mediator model for providing integrated and unified access to multiple taxonomy-based sources. Each source comprises a taxonomy and a database that indexes objects under the terms of the taxonomy. A mediator comprises a taxonomy and a set of relations between the mediator’s and the sources’ terms, called articulations. By combining different modes of query evaluation at the sources and the mediator and different types of query translation, a flexible, efficient scheme of mediator operation is obtained that can accommodate various application needs and levels of answer quality. We adopt a simple conceptual modeling approach (taxonomy and inter-taxonomy mappings) and we illustrate its advantages in terms of ease of use, uniformity, scalability, and efficiency. These characteristics make this proposal appropriate for a large-scale network of sources and mediators.

Keywords: Mediators – Taxonomies – Approximate query translation – Information integration

1 Introduction

The need for integrated and unified access to multiple information sources has stimulated research on mediators (initially proposed in [78]). Roughly, a mediator is a secondary information source aiming at providing a uniform interface to a number of underlying sources (which may be primary or secondary). Users submit queries to the mediator. Upon receiving a user query, the mediator queries the underlying sources. This involves selecting the sources to be queried and formulating the query to be sent to each source. These tasks are accomplished based on what the mediator “knows” about the underlying sources. Finally, the mediator appropriately combines the returned results and delivers the final answer to the user.

In this paper we consider information sources over a common domain consisting of a denumerable set of objects. For example, in the environment of the Web, the domain could be the set of all Web pages, specifically, the set of all pointers to Web pages. Each source has a taxonomy, i.e., a structured set of names, or terms, that are familiar to the users of the source. In particular, the taxonomies considered in this paper consist of a set of terms structured by a subsumption relation. In addition, each source maintains a database storing objects that are of interest to its users. Specifically, each object in the database of a source is indexed under one or more terms of the taxonomy of that source. In quest for objects of interest, a user can browse the source taxonomy until he reaches the desired terms, or he can query the source by submitting a boolean expression of terms. The source will then return the appropriate set of objects. In the environment of the Web, general-purpose catalogs, such as Yahoo! or Open Directory,1 domain-specific catalogs/gateways (e.g., for medicine, physics, tourism), as well as personal bookmarks of Web browsers can be considered as examples of such sources.

However, although several sources may carry information about the same domain, they usually employ different taxonomies, with terms that correspond to different natural languages, or different levels of granularity. For example, consider two sources S1 and S2 that both provide access to electronic products as shown in Figs. 1a and 1b. Each source consists of a taxonomy plus a database that indexes objects under the terms of that taxonomy. However, the two sources provide different information about electronic products, as seen in the figures. Suppose now that we want to provide unified access to these two sources through a single taxonomy that is familiar to a specific group of users. An example of such a unifying taxonomy is shown in Fig. 1c and constitutes part of what we call a “mediator”.

A mediator is a secondary source that can bridge the heterogeneities that may exist between two or more sources in order to provide unified access to those sources. Specifically, a mediator has a taxonomy with terminology and structuring that reflects the needs of its potential users but does not maintain a database of objects. Instead, the mediator maintains a number of articulations to the sources. An articulation to a source is a set of relationships between the terms of the mediator and the terms of that source. These relationships are defined by the designer of the mediator at design time and are

1 http://dmoz.org
stored at the mediator. Figure 2 shows the general architecture of a mediator.

Users formulate queries over the taxonomy of the mediator, and it is the task of the mediator to choose the sources to be queried and to formulate the query to be sent to each source. To this end, the mediator uses the articulations to transcribe queries over its own taxonomy to queries over the taxonomies of the articulated sources. Then it is again the task of the mediator to combine the results returned by the sources appropriately to produce the final answer.

An essential feature that distinguishes our work is that we adopt a simple conceptual modeling approach for both sources and mediators. This conceptual modeling approach has the following advantages: (a) it is very easy to create the conceptual model of a source or a mediator, and (b) the integration of information from multiple sources can be done very easily. Indeed, as we shall see, the articulations offer a uniform and easy-to-use method to bridge naming, contextual, and granularity heterogeneities between the conceptual models of the sources. Given this conceptual modeling approach, the mediator does not have to tackle complex structural differences between the sources (as happens in mediators for relational databases).

Another essential feature that distinguishes our approach is that a source can provide two types of answer to a given query, namely, a sure answer or a possible answer. The first type of answer is appropriate for a user who does not want to retrieve objects that are not relevant to his information need, while the second is for a user who does not want to miss objects that are relevant to his information need. Moreover, as exact translation of user queries is not always possible, a user query to the mediator admits two types of approximation—lower or upper translation.

What kind of translation will be used at the mediator level and what kind of answer will be requested at the source level is decided by the mediator designer at design time and/or the mediator user at query time. Therefore, a prominent feature of our approach is that sources and mediators can operate in a variety of modes according to specific application needs. As a consequence, our mediators are quite flexible and can adapt to a variety of situations.

A main objective of this paper is to prescribe easy-to-use and formally sound methods for building mediators. In the context of the Web, our mediators can be used for providing unified access to multiple Web catalogs. An advantage of our approach is that a mediator can be constructed quite easily; therefore ordinary Web users can use it to define their own mediators. In this sense, this approach can be used for personalizing existing Web catalogs. Furthermore, it can be used for building mediators over XFML [1] information bases (XFML aims at applying the faceted classification paradigm in the context of the Web).

The remainder of the paper is organized as follows. Section 2 describes the information sources and the query answering process at a single source. Section 3 defines the architecture of a mediator over a set of sources and the different modes in which a mediator can operate. Section 4 discusses query evaluation, and Sect. 5 discusses enhancements of the query answering process. Section 6 discusses various extensions of our model. Section 7 discusses related work and, finally, Sect. 8 concludes the paper and discusses further research. All proofs are given in the Appendix.

2 The sources

Why taxonomies

Taxonomies are probably the oldest and most widely used conceptual modeling tool. Nevertheless, it is a powerful tool still used in Web directories (e.g., Google and Yahoo!), content management (hierarchical structures are used to classify documents), Web publishing (many authoring tools require one to organize the contents of portals according to some hierarchical structure), Web services (services are typically classified in a hierarchical form), marketplaces (goods are classified in hierarchical catalogs), personal file systems, personal bookmarks for the Web, libraries (e.g., Thesauri [40]), and in very large collections of objects (e.g., see [61]). Although more sophisticated conceptual models (including concepts, attributes, relations, and axioms) have emerged and have recently been employed even for metatagging on the Web [48,75], almost all of them have a backbone consisting of a subsumption hierarchy, i.e., a taxonomy.

Furthermore, a taxonomy-based conceptual modeling approach has several advantages in large and open domains. In a very broad domain, such as the set of all Web pages, it is not easy to identify the classes of the domain because the domain is too wide and different users, or applications, conceptualize it differently, e.g., one class of the conceptual model according to one user may correspond to a value of an attribute of a class of the conceptual model according to another user. For example, Fig. 3 shows two different conceptual models for the same domain. We consider only two objects of the domain, denoted by the natural numbers 1 and 2.

The conceptual model of Fig. 3a is appropriate for building an information system for a furniture store, while the conceptual model of Fig. 3b is appropriate for building an information...