

A Mediator for Heterogeneous Gazetteers

Alexandre Gazola, Daniela F. Brauner, Marco A. Casanova

Departamento de Informática – PUC-Rio
Rua Marquês de São Vicente, 225 – 22.453-900 – Rio de Janeiro – RJ – Brazil
{agazola, dani, casanova}@inf.puc-rio.br

***Abstract.** Gazetteers are catalogs of geographic objects, typically classified using terms taken from a thesaurus. Mediated access to several gazetteers requires the use of a strategy to deal with the heterogeneity of different thesauri. This paper outlines the implementation of a mediator for heterogeneous gazetteers. The mediator incorporates an instance-based technique to align thesauri that uses the results of user queries as evidences.*

1. Introduction

One of the challenges the database community has been addressing for quite some time now is to define mediators that give users a unique, homogeneous and uniform view of a set of heterogeneous, autonomous data sources [Barbosa et al. 2004]. At the heart of this challenge lies the problem of matching the conceptual schemas that describe the data sources. [Rahm and Bernstein 2001] survey early attempts to automate schema matching, mostly exploring syntactical similarities to match schema elements. [Brauner et al. 2006] provide an example of a different technique, based on an online mapping estimator that gradually creates weighted relationships between terms of distinct thesauri by post-processing the result sets that the user queries return.

In this paper, we describe the design of a mediator for heterogeneous gazetteers. The mediator allows users to search for objects in registered gazetteers, using terms from a preferred thesaurus or using keywords. The major contribution of the paper is a concrete implementation of the technique described in [Brauner et al. 2006] to map terms from distinct thesauri. In short, a gazetteer is a data dictionary that contains a list of geographic names along with their spatial representations and other descriptive information [Hill et al. 1999]. Also, gazetteers typically classify geographic objects using a thesaurus, which is a list of structured terms that standardizes the vocabulary used for indexing [Unesco 1995].

The remainder of this paper is organized as follows. Section 2 describes the mediator architecture. Section 3 outlines the prototype implementation. Finally, Section 4 contains the conclusions and directions for future work.

2. Mediator Architecture

Figure 1 illustrates the architecture we propose for a mediator for heterogeneous gazetteers.

The architecture supports the thesauri mapping technique described in [Brauner et al. 2006]. Briefly, given two catalogs, C_A and C_B , with thesauri T_A and T_B , if a query returns an object c from C_A classified as t_A (a term of T_A) and, again, c from C_B , but

classified as t_B (a term of T_B), then c establishes some evidence that t_A and t_B map to each other. This technique relies on the assumption that the mediator is able to recognize whether data from different catalogs represent the same object or not. For instance, in geographic information applications, the mediator may use the spatial location of an object, if both catalogs store such information. The mediator then operates based on the runtime generation of evidences, and not on precisely defined thesauri alignments, which are hard to obtain a priori.

Returning to Figure 1, the User Interface Module (UIM) is responsible for the communication between the user and the mediator. The UIM accepts user queries and returns their answers. It communicates with the Query Manager Module (QMM). The QMM is responsible for decomposing user queries into queries over the data sources and for submitting such queries to the registered data sources. The QMM communicates with the *Wrappers* layer to access remote and local sources. During the query decomposition process, the QMM may need to communicate with the local sources (Local Wrappers Module - LWM) to retrieve existing mappings to formulate queries over the data sources. Moreover, the QMM communicates with the LWM to store query responses in the local cache database. The Mapping Rate Estimator Module (MREM) is an autonomous module that is responsible for accessing the local cache database to compute the mapping rates for the thesauri terms.

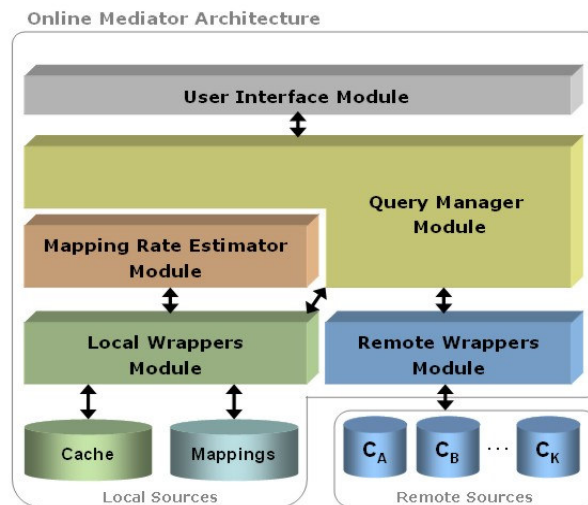


Figure 1. The Mediator Architecture

3. Mediator Prototype Implementation

Using the technique presented in [Brauner et al. 2006] and a simplification of the architecture described in Section 2, we implemented a mediator that handles queries submitted to the Alexandria Digital Library (ADL) gazetteer and to the GEOnet Names Server (GNS).

The ADL gazetteer adopts the ADL Feature Type Thesaurus and implements a specific protocol [Janée and Hill 2002] that provides online access to the ADL gazetteer data. GNS is a gazetteer that stores millions of entries about geographic objects

SBBD 2007
I Sessão de Pôsteres

worldwide. GNS provides access to the database of foreign geographic feature names for the National Geospatial-Intelligence Agency and the U.S. Board on Geographic Names [GNS 2006].

The mediator was implemented in Java as a Web application, and the gazetteer data were kept in a local database to simplify the prototype implementation. We intend to implement remote access to the gazetteers, using a common interface, such as the Gazetteer Services Profile of the WFS interface, defined by the Open Geospatial Consortium [OGC 2006].

The mediator supports two kinds of queries: by *classification* and by *keyword*. In a query by classification, the user selects a term *t* for a thesaurus of his choice, among those registered with the mediator. The mediator uses the mappings already discovered to map *t* into terms of the other thesauri. If no mapping is found, the mediator asks the user to manually inform the corresponding terms from the other thesauri. The query is processed and its results are analyzed to update the mapping rate estimations, which will be used in the next queries. In a query by keyword, the user types a keyword that the mediator uses to search the name field of the registered gazetteers. Then, each classification term of the returned objects are processed against each other, again to update the mapping rate estimations.

The mediator assumes that it is possible to identify which entries from different gazetteers represent the same geographic object. In our implementation, we used the latitude and longitude of the centroid of the object to detect equivalent entries.

As an example, Table 1 displays the mapping rates calculated by the mediator after processing several queries. Focusing on the first line, it reflects one query by the keyword “Rio”, and one query by classification using the term “populated places” (a term of the ADL thesaurus). Now, we observe that ADL has 22.269 items classified as “populated places”, whereas GNS has 43.617 items classified as “PPL” (the term of the GNS thesaurus equivalent to “populated places”). After analyzing the results of the first query, the mediator computed the mapping rate between “populated places” and “PPL” as 0.23. After the second query, the mapping rate was updated to 0.98, which suffices to establish that “populated places” and “PPL” map to each other. Indeed, the empirical analysis carried out in [Brauner et al. 2006] showed that 0.4 is a reasonable threshold to establish thesauri mappings with great accuracy.

Table 1. Example of the mapping rate estimations generated by the mediator

Base Term	Target Term	Mapping Rate	Number of Queries
populated places (ADL)	PPL (GNS)	0.986027177140462	2
streams (ADL)	STM (GNS)	0.909090909090909	1
reservoirs (ADL)	RSV (GNS)	0.5	1
administrative areas (ADL)	ADM1 (GNS)	0.5	1
railroad features (ADL)	STM (GNS)	0.25	1
railroad features (ADL)	RSTP (GNS)	0.142857142857143	1
railroad features (ADL)	RSTN (GNS)	0.1	1

4. Conclusions and future work

In this paper, we outlined the implementation of a mediator for gazetteers. The mediator was used to validate the technique proposed in [Brauner et al. 2006], which adopts an estimator that gradually creates and updates weighted relationships between terms of distinct thesauri by post-processing common instances from two gazetteers.

This is the preliminary step towards the development of an instance-based schema matching heuristic to enable mediated access to heterogeneous catalogs. More specifically, we will use the strategy implemented in this paper to align thesauri terms, and a modification of the strategy described in [Wang et al. 2004] to align the sets of attributes of the catalog schemas. The strategy is to adopt a global schema along with a small sample of typical instances pertaining to the domain of discourse that will help determine the schema alignments.

Acknowledgments

This work is partly supported by CNPq under grants 550320/02-1 and 140417/05-2, and Faperj under grant E-26/100.128/2007.

References

- Brauner, D. F., Casanova, M. A. and Milidiú, R. L. (2006) “Mediation as Recommendation: An Approach to Design Mediators for Object Catalogs”. Proc. 5th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2006). Montpellier, France.
- GNS (2006), “GEOnet Names Server”, U.S. National Geospatial-Intelligence Agency, USA. Available at: <http://gnswww.nga.mil/geonames/GNS>.
- Hill, L. L., Frew, J., and Zheng, Q. (1999), “Geographic names: The implementation of a gazetteer in a georeferenced digital library.” D-Lib (January 1999). Available at: <http://www.dlib.org/dlib/january99/hill/01hill.html>
- Janée, G., and Hill, L. L. (2002), “ADL Gazetteer Protocol. (Version 1.1)”. Alexandria Digital Library Project. Retrieved May 2 2003. Available at <http://www.alexandria.ucsb.edu/gazetteer/protocol/>
- OGC (2006), “Gazetteer Service Profile of the Web Feature Service Implementation Specification”. Available at: <http://www.opengeospatial.org>
- UNESCO (1995). “UNESCO Thesaurus”, United Nations Educational, Scientific and Cultural Organization. Available at <http://www.ulcc.ac.uk/unesco/index.htm>
- Wang, J., Wen, J.-R., Lochovsky, F., and Ma, W.-Y. (2004) “Instance-based schema matching for web databases by domain-specific query probing”. In Proceedings of the 30th VLDB Conference, Toronto, Canada.