

Mediation as Recommendation: An Approach to Design Mediators for Object Catalogs¹

Daniela F. Brauner, Marco A. Casanova, Ruy L. Milidiú

Department of Informatics – PUC-Rio

Rua Marquês de S. Vicente, 225 – Rio de Janeiro, RJ – Brazil
{dani, casanova, milidiu}@inf.puc-rio.br

Abstract. A catalog holds information about a set of objects, typically classified using terms taken from a given thesaurus. Mediated access to a collection of catalogs over the same domain therefore requires some strategy to deal with multiple thesauri, which represent different classifications for the same domain. This paper proposes an approach using online mapping rate estimations to define weighted relationships between terms of distinct thesauri. The mediator then uses such relationships to remap keyword-based queries to the different catalogs. Moreover, query answers provide valuable feedback to adjust the relationship weights, thereby improving the mediator accuracy.

1. Introduction

A *catalog* is a database that stores information about a set of objects, classified using terms taken from a given *object type thesaurus*. The design of a mediator for a collection of catalogs therefore requires aligning distinct object type thesauri, which is the central question we address in this paper.

We address this question by designing an online mapping rate estimator that gradually creates weighted relationships between terms of distinct thesauri by post-processing result sets returned by user queries. 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 $t_a \in T_A$ and, again, c from C_B , but classified as $t_b \in T_B$, then c establishes some evidence that t_B maps into t_a . Note that this strategy depends on the assumption that the mediator can recognize when data from different catalogs represent the same object or not. For instance, in e-commerce applications the mediator may use the manufacture's part numbers, if both catalogs store such information. Likewise, the mediator may use the object's spatial location in geographic information applications to try to deduce that two objects are indeed the same. The mediator then operates based on the online generation of evidences, and not on precisely defined thesauri alignments, which are very difficult to define a priori.

¹ This work is partially supported by CNPq under grants 550250/05-0, 140417/05-2 and 552068/02-0.

2. Estimating Relationships between Terms of two Thesauri

For simplicity, we assume that we have just two catalogs, C_A and C_B , storing objects from the same domain, classified using thesauri T_A and T_B , respectively. Also, we will be interested in mapping terms from T_A into T_B . However, the discussion can be generalized to bi-directional mappings and to more than two catalogs.

We say that entries $c_a \in C_A$ and $c_b \in C_B$ are *equivalent*, denoted $c_a \equiv c_b$, when they represent the same (real-world) object. The exact procedure that computes instance equivalence depends on the application.

A *user section* is a pair of queries Q_A over C_A and Q_B over C_B , submitted through the mediator. We assume that a user section contains queries that try to retrieve objects from C_A and C_B that are similarly classified. After the training process, the mediator will be able to recommend how to query C_B based on the terms used to query C_A .

Let $t_a \in T_A$ and $t_b \in T_B$ in what follows. The mediator maintains $P(t_a, t_b)$, the *mapping rate estimator* for t_a and t_b , which estimates the frequency that the term t_a maps to the term t_b . The mediator computes $P(t_a, t_b)$ as follows.

The mediator stores $n(t_a, t_b)$, the sum of the all occurrences of pairs of objects $c_a \in C_A$ and $c_b \in C_B$ such that: (1) $c_a \equiv c_b$; (2) the types of c_a and c_b are t_a and t_b , respectively; (3) c_a and c_b were observed in a previous user section. The mediator also stores $n(t_a)$, the sum of the all occurrences of objects $c_a \in C_A$ such that: (1) the type of c_a is t_a ; (2) c_a was observed in a previous user section.

The mediator post-processes the result sets a new user section and computes $\Delta n(t_a, t_b)$ and $\Delta n(t_a)$, defined exactly as $n(t_a, t_b)$ and $n(t_a)$, except that the objects are those in the result sets of the new user section. Then, the mediator recomputes $P(t_a, t_b)$ as follows:

$$P(t_a, t_b) = \frac{\Delta n(t_a, t_b) + \alpha \cdot (n(t_a, t_b) + \Psi)}{\Delta n(t_a) + \alpha \cdot (n(t_a) + 1)}$$

where

- α is a coefficient that takes values from the set $\{0.01, 0.1, 0, 1, 10, 100\}$, calibrated during the model validation process
- $\Psi = \frac{1}{|T_B|}$ is a smoothing coefficient assumed as the inverse of the size of the thesaurus of the second term.

Note that the above equation is symmetric in t_a and t_b and can be easily adapted to compute estimations for the frequency that the terms in T_B map into terms in T_A .

To validate and test the estimation model, we used the ADL Gazetteer (<http://www.alexandria.ucsb.edu/gazetteer>) and the GEOnet Names Server (<http://gnswww.nga.mil/geonames/GNS>). We stored data from both gazetteers locally and partitioned the data into a tune set and a test set. We applied the 6-fold cross-validation technique to calibrate the model parameters. As a result, we obtained 26 pairs of terms from T_A to T_B aligned with mapping rate greater than 0.4, with accuracy of 89.7% and recall 81.3%. From T_B to T_A , we obtained 44 pairs of terms aligned with accuracy of 93.6% and recall of 95.7%.