Towards Gazetteer Integration
Through an Instance-based Thesauri Mapping Approach

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

Departamento de Informática – PUC-Rio
Rua Marquês de São Vicente, 225 – 22.453-900 – Rio de Janeiro – RJ – Brazil

`{dani, casanova, milidiu}@inf.puc-rio.br`

Abstract. *Gazetteers are catalogs of geographic features, typically classified using a feature type thesaurus. Integrating gazetteers is an issue that requires some strategy to deal with multiple thesauri, which represent different classifications for the geographic domain. This paper proposes an instance-based approach to define mapping rates between terms of distinct feature type thesauri in order to enable the reclassification of the data migrated from one gazetteer to another.*

1. Introduction

A *gazetteer* is a database that stores information about a set of geographic features, classified using terms taken from a given *feature type thesaurus*. Gazetteers could be used as information sources of annotation systems of geographic data [Leme 2006]. An annotation system could use many different gazetteers to get information about the data to be catalogued. However, as in a data-warehouse creation process, gazetteer integration requires aligning feature type thesauri, which is the central question we address in this paper.

Our approach uses a mapping rate estimator that estimates weighted relationships between terms of distinct thesauri by pre-processing common instances from two gazetteers. Let G and G' using thesauri T and T', respectively, be the gazetteers to be integrated. Quite simply, if we have data about a geographic feature f from G classified as t (a term from T) and, again, data about f from G', but classified as t' (a term from T'), then f establishes some evidence that t' maps into t. Note that this strategy depends on the assumption that we can recognize when data from G and G' represent the same geographic feature or not. In this paper, we use the feature's spatial location from G and G', to deduce that a common set of data from G and G' indeed represent the same geographic features or not.

As for related work, in the area of mediator construction, we may single out the OBSERVER system [Mena et al. 1996; Mena et al. 2000], which uses multiple ontologies, described in a Description Logics formalism, to access heterogeneous and distributed data sources. OBSERVER requires that conventional mappings between a data source and the base ontology be manually defined. By contrast, our approach automatically generates weighted mappings, working with thesauri.

In the area of ontology mapping, we may highlight the GLUE system, that makes use of multiple learning strategies to help find mappings between two ontologies [Doan et al. 2003], the AnchorPROMPT ontology alignment tool, that automatically

identifies semantically similar terms [Noy et al. 2003], and the Chimaera environment, that provides a tool to merge ontologies based on their structural relationships [McGuineess et al. 2000]. These three tools work with fully formalized ontologies and, to a varying extent, depend on user intervention. The CATO tool aligns thesauri using mostly syntactical similarities between terms and the thesauri structure [Breitman et al. 2005].

Our approach differs from such systems in two aspects. First, like CATO, we work only with the terms and their structure (the broader term/narrow term relationship). That is, we do not require a fully formalized terminology, using an ontology language. However, unlike CATO, to align two terms, we draw evidence from the way the gazetteers classify geographic features, not merely from a syntactical similarity between the terms.

Castano et al. (2004) describe the H-Match algorithm to dynamically match ontologies. H-Match provides, for each concept from an ontology, a ranked list of similar concepts in the other ontology. Four matching models are used to dynamically adjust the matching process to different levels of richness of the ontology descriptions. Spertus et al. (2005) evaluate the performance of six similarity measures, used to recommend related communities to members of Orkut social network communities, adopting the L2 vector normalization (L2-Norm) measure.

This paper is organized as follows. Section 2 summarizes preliminary definitions. Section 3 contains a motivating example. Section 4 describes our instance-based approach to thesauri mapping, including experimental results. Finally, section 5 contains the conclusions and directions for future work.

2. Gazetteers and Thesauri

A *thesaurus* is defined as *"a structured and defined list of terms which standardizes words used for indexing"* [UNESCO 1995] or, equivalently, *"the vocabulary of a controlled indexing language, formally organized so that a priori relationships between concepts (for example as "broader" and "narrower") are made explicit"* [ISO-2788 1986]. A thesaurus usually provides: a *preferred term,* defined as the term used consistently to represent a given concept; a *non-preferred term*, defined as the synonym or quasi-synonym of a preferred term; relationships between the terms, such as *narrower term (NT)*, indicating that a term – the *narrower term* – refers to a concept which has a more specific meaning than another term – the *broader term (BT)*.

A *gazetteer* is *"a geographical dictionary (as at the back of an atlas) containing a list of geographic names, together with their geographic locations and other descriptive information"* [Wordnet 2005]. For our purposes and omitting details, we consider that a gazetteer is a geographic object catalog, where each object has as attributes:

− a unique *object ID*;

− a unique *object type*, whose value is a term taken from an *object type thesaurus*;

− a *name*, which takes a character string as value;

− optionally, a *location*, which approximates the object's position on the Earth's surface.

210

For simplicity, we assume that the object type is unique, and that each object has only one name (which is not necessarily a key). We note that geographic objects are often called *geographic features*, or simply *features* [Percivall 2003]. Hence, a gazetteer thesaurus is also referred to as a *feature type thesaurus*.

Let *G* be a gazetteer with thesaurus *T* and *G'* be a gazetteer with thesaurus *T'*. Supposes that one wants to load the data from *G'* into *G*, the question is how to remap the feature types from thesaurus *T'* into *T*.

Assuming that the gazetteers are homogeneous, i.e., given any two features, *f* and *f'*, from any two gazetteers *G* and *G'*, it is possible to detect when *f* and *f'* denote the same real world object. This is more an assumption than a definition since we leave it open what is the exact procedure used to detect identical objects. We are interested in be able to reclassify features from *G'* using feature type thesaurus from *G*, i.e., remap feature types from *T'* into feature types from *T*.

3. A Motivating Example

As a motivating example, we will use OpenCyc and two gazetteers that are available over the Web, the GEOnet Names Server and the Alexandria Digital Library Gazetteer.

OpenCyc [Cyc 2005] is an upper ontology describing approximately 47,000 concepts and 306,000 assertions. We will be mostly interested in the Cyc knowledge base instances that describe cities and countries. The GEOnet Names Server (GNS) [GNS 2006] provides access to the National Geospatial-Intelligence Agency (NGA) and the U.S. BGN database of foreign geographic names, containing about 4 million features with 5.5 million names. The Alexandria Digital Library (ADL) Project [ADL 1999; Hill et al. 1999] is a research program to model, prototype and evaluate digital library architectures, gazetteer applications, educational applications, and software components. The ADL Gazetteer has approximately 5.9 million geographic names, classified according to the ADL Feature Type Thesaurus (FTT).

Figures 1(a) and 1(b) show fragments of the ADL Feature Type Thesaurus and of the OpenCyc thesaurus. We note that the GEOnet Names Server classification scheme does not have a formal hierarchical organization. However, Table 1 shows the fragment of the GEOnet classification scheme equivalent to those in Figure 1.
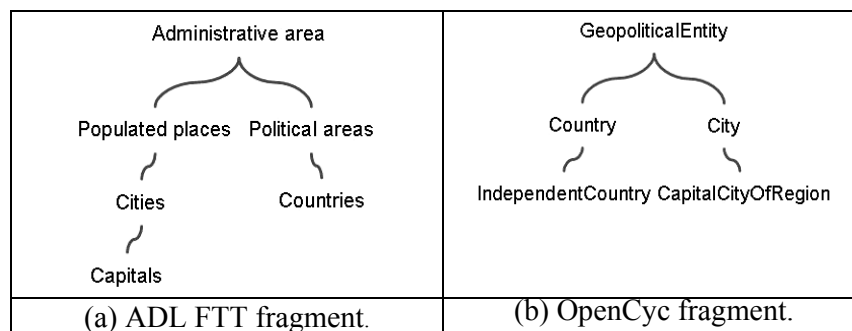

| (a) ADL FTT fragment. | (b) OpenCyc fragment. |

Figure 1. Fragments of the ADL and OpenCyc feature type thesauri.

Table 1. Fragment of the GEOnet classification scheme.

Code	Description Text
PCLI	"Independent political entity"
AREA	"A tract of land without homogeneous character or boundaries"
PPL	"Populated place"
PPLA	"Seat of a first-order administrative division"
PPLC	"Capital of a political entity"
PCLI	"Independent political entity"

In what follows, we will refer to the ADL Gazetteer, GEOnet Names Server and OpenCyc, respectively, as G_A, G_B, G_C, and to their thesauri as T_A, T_B, T_C. We will consider only countries and cities in the examples that follow. For simplicity, we assume that the name (in English) uniquely identifies a country in all three gazetteers; similarly, the city name, together with the name of the upper level administrative division, uniquely identifies a city in all three gazetteers.

We will illustrate how to integrate gazetteers using an instance-based technique that aims at mapping *feature type thesauri* to enable the reclassification of the instances migrated from one gazetteer to another.

To illustrate thesauri differences, Table 2 shows examples of information about how the gazetteers classify entries, collected from queries that searched the three gazetteers for the countries and cities listed in the first column. For example, if the user accesses the ADL Gazetteer to obtain information about 'Brazil', the answer will indicate that the ADL Gazetteer classifies 'Brazil' as 'Countries'; if he then access OpenCyc for 'Brazil', the answer shows that OpenCyc classifies 'Brazil' as 'IndependentCountry'; likewise, GEOnet classifies 'Brazil' as 'PCLI'. In fact, all 5 entries in Table 2 that the ADL Gazetteer classifies as 'Countries', OpenCyc classifies as 'IndependentCountry' and GEOnet as 'PCLI'. Hence, we have evidences that these three terms map to each other. Therefore, if we would like to load G_B into G_A, this small sample provides us with an evidence that instances from G_B classified as 'PCLI' from T_B have to be loaded to G_A reclassified as 'Countries' from T_A. Moreover, it does not detect any conflicting classifications in this small sample.

Table 2. Results of querying countries and cities in the ADL Gazetteer, the GEOnet Names Server and OpenCyc.

Entry name	ADL Gazetteer (T_A)	GEOnet (T_B)	OpenCyc (T_C)
Brazil	Countries	PCLI	IndependentCountry
Canada	Countries	PCLI	IndependentCountry
Germany	Countries	PCLI	IndependentCountry
Italy	Countries	PCLI	IndependentCountry
Belgium	Countries	PCLI	IndependentCountry
Scotland – UK	AdministrativeArea	AREA	Country
Wales – UK	AdministrativeArea	AREA	Country
Rio de Janeiro – Brazil	Populated Places	PPLA	City
São Paulo – Brazil	Populated Places	PPL	City
Rome – Italy	Capitals	PPLC	CapitalCityOfRegion
Brussels – Belgium	Capitals	PPLC	CapitalCityOfRegion

4. Instance-based Thesauri Mapping Approach

4.1 Conceptual Model

Our goal is to enable the integration of gazetteers, knowing that they may use different thesauri to classify their features. Similarly to a data-warehouse creation process, we are loading data from one source to another and dealing with the heterogeneities of the vocabularies used to classify source objects. To solve vocabulary conflicts, we focus on estimating weighted relationships between concepts of distinct thesauri. To achieve this goal, we propose to collect statistics about the common instances from both gazetteer instances.

Suppose that we have two gazetteers, G_A and G_B, classified using thesauri T_A and T_B, respectively. Suppose also that we are interested in mapping terms from T_A to T_B.

We say that features f_a in G_A and f_b in G_B, respectively, are *equivalent*, denoted $f_a \equiv f_b$, when they represent the same (real-world) object; in this case, we also say that t_a and t_b *map to each other*, where t_a and t_b are the types of f_a and f_b, respectively. The exact procedure that computes instance equivalence depends on the application, as previously discussed.

We define F_A as the set of all $f_a \in G_A$ such that there is $f_b \in G_B$ such that $f_a \equiv f_b$ (and similarly for $F_B \subseteq G_B$).

For each pair of terms $t_a \in T_A$ and $t_b \in T_B$, we compute the following information:

- n(t_a,t_b), the sum of the occurrences of pairs of objects f_a and f_b such that:
  - $f_a \in G_A$ and $f_b \in G_B$
  - $f_a \equiv f_b$
  - t_a and t_b are the types of f_a, and f_b, respectively

- P(t_a,t_b), an estimation for the frequency that the term t_a maps to the term t_b, for each pair of terms $t_a \in T_A$ and $t_b \in T_B$. We call P(t_a,t_b) the *mapping rate estimator* for t_a and t_b.

For each term $t_a \in T_A$, we store the following information:

- n(t_a), the sum of the occurrences of objects $f_a \in F_A$ such that t_a is the type of f_a

Returning to the example in Section 3, suppose that the user wants to load the GEOnet Names Server (G_B) into the ADL Gazetteer (G_A). We will compute n(t_a), n(t_a,t_b) and P(t_a,t_b), as explained in Section 4.2.

In the geographic information systems domain, we have various geo-referencing schemes that associate each geographic object with a description of its location on the Earth's surface. This location acts as a universal identifier for the object, or at least an approximation thereof. In our approach, we use the object's location to detect equivalent instances and to count the frequency of pairs of terms from different gazetteer thesauri. In other words, we analyze which entries from different gazetteers represent the same geographic object and then calculate a similarity measure between their respective types.

4.2 Statistical Model

This section outlines the statistical model we adopt. Assume that we have already computed F_A and F_B. We use F_A and F_B to estimate $n(t_a)$, $n(t_a,t_b)$ and $P(t_a,t_b)$ as follows:

1. Compare the features in F_A with those in F_B to count the pairs of objects $f_a \in F_A$ and $f_b \in F_B$ such that $f_a \equiv f_b$.

2. Compute $n(t_a,t_b)$, the number of occurrences of pairs of objects $f_a \in F_A$ and $f_b \in F_B$ such that $f_a \equiv f_b$ and t_a and t_b are the types of f_a, and f_b respectively.

3. Compute $n(t_a)$, the number of occurrences of t_a in F_A.

4. Compute $P(t_a,t_b)$ using equation (1):

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

where

$\Delta = \dfrac{1}{|T_B|}$ is a smoothing coefficient assumed as the product of 1 and the number of terms of the thesaurus T_B;

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

4.3 Experiments with Geographic Data

In order to illustrate the mapping rate estimation model proposed in Section 4.2, we present results using real feature type thesauri in the GIS domain. Section 4.3.1 describes how the data was obtained from the ADL Gazetteer (G_A) and the GEOnet Names Server (G_B). Section 4.3.2 discusses how the model was validated and calibrated. Section 4.3.3 contains the test results.

4.3.1 Data Collection

To facilitate the training step, data were collected from remote gazetteers servers and stored locally. G_A was consulted using version 1.2 of the ADL Gazetteer Service Protocol, an XML- and HTTP-based protocol for accessing the ADL Gazetteer [Janée and Hill 2004]. Several queries where submitted to G_A, restricted to the Brazilian geographic area, retrieving 16,783 registries in the standard ADL report format (in XML). The returned XML was parsed and the registries were stored in a relational database. As for G_B, data were downloaded from the GEOnet Names Server Website, which contains files with information about geographic names, covering countries or geopolitical areas. The files are not in the GEOnet gazetteer format, but in a special format amenable to input to a database. The downloaded Brazilian file had 87,608 registries. The available data were partitioned into a tuning set and a testing set, used to tune and to test the model, respectively.

Table 3. ADL Feature Type Thesaurus relationships.

Abbreviation	Relationship Name
USE	Use
UF	Used for
USW	Used with
UFW	Used for with
BT	Broader term
NT	Narrower term
RT	Related term
SN	Scope note
DF	Definition
HN	History note

Table 4. Top terms from ADL FTT and GEOnet thesaurus.

ADL FTT top terms	GEOnet thesaurus top terms
Administrative Areas	Populated Place
Hydrographic Features	Administrative Region
Land Parcels	Area
Manmade Features	Vegetation
Physiographic Features	Streets/Highways/Roads
Regions	Hypsographic
	Hydrographic
	Undersea
	Spot Features

Moreover, thesauri data were collected and also stored locally. The ADL Feature Type Thesaurus (FTT) has 1,262 terms, organized hierarchically and related using an extended set of the basic thesaurus relationships as presented in Table 3. An example including the list of the ADL FTT top terms is presented in Table 4. For this experiment, we consider that the size of T_A is the number of preferred terms (210 terms). The preferred terms are the terms used to classify objects, whereas the other terms are related to the preferred terms through the relationships USE, UF, UFW and USW. The GEOnet thesaurus (T_B) has 642 terms, organized under a single category level including 9 top terms (Table 4). The GEOnet thesaurus includes the term code, name and a textual description.

4.3.2 Model Evaluation

To validate the mapping rate estimation model, the data collected was partitioned into seven disjoint datasets. Six of the datasets were used as tuning sets, and one as the testing set.

The tuning set was partitioned into six training sets and six validation sets to apply the 6-fold cross-validation method to estimate the accuracy of the model. Table 5 shows the six tuning sets and its training (T_k) and validation (V_k) sets with the number of term pairs covered. The validation sets (V_k) were manually labeled with True or False for each occurrence of pairs of terms. Pairs labeled with True indicate that the terms indeed map to each other. The labeling was made by comparing thesauri descriptions and a brief check of equivalent entries, with the help of a geographic domain expert.

Table 5: Tuning sets for 6-fold cross-validation technique.

Id	Dataset.Id	Dataset	Pairs
Tn1	T_1	Ex1	92
	V_1	Ex2, Ex3, Ex4, Ex5, Ex6	180
Tn2	T_2	Ex2	87
	V_2	Ex1, Ex3, Ex4, Ex5, Ex6	189
Tn3	T_3	Ex3	67
	V_3	Ex1, Ex2, Ex4, Ex5, Ex6	197
Tn4	T_4	Ex4	46
	V_4	Ex1, Ex2, Ex3, Ex5, Ex6	191
Tn5	T_5	Ex5	68
	V_5	Ex1, Ex2, Ex3, Ex4, Ex6	183
Tn6	T_6	Ex6	78
	V_6	Ex1, Ex2, Ex3, Ex4, Ex5	174

In the k-fold cross-validation method the model is trained and tested k times; each time it is trained with T_k and validated with V_k. The cross-validation estimate of accuracy is the overall number of pairs that were correctly matched (with respect to the validation sets), divided by the number of previously labeled pairs.

The *threshold mapping rate* is the value above which the mapping rates $P(t_a,t_b)$ are considered. It was estimated as follows. The mapping rates of the pairs of terms of the training sets were estimated several times, varying the threshold value from 0 to 1, by 0.1, to discover the best value (see Figure 3). The cross-validation process compares these results with the labeled pairs from each validation set. Figure 3 shows that better results were obtained with threshold mapping rate equal to 0.4 (to the cross validation from T_A to T_B).
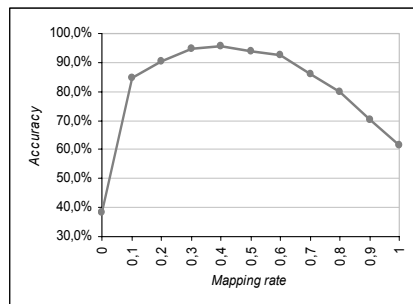
Figure 3. 6-fold cross-validation results from T_A to T_B.

4.3.3 Test

To test the mapping rate estimation model we use the testing set and the mapping rate 0.4 estimated during the model evaluation.

As a result of the test step, we have 26 pairs of terms aligned with mapping rate greater then 0.4 from T_A to T_B with accuracy of 89.7% and recall 81.3%.

Table 6. Aligned terms during test step.

t_a	t_b	$P(t_a, t_b)$
agricultural sites	FRM	0.96974
bays	BAY	0.50039
forests	RESF	0.50078
islands	ISL	0.93422
lakes	LK	0.91849

Table 6 shows examples of the aligned terms. For example, 'agricultural sites' from T_A aligns with 'FRM' from T_B with mapping rate '0.96974'. These values indicate that features migrated from G_B into G_A, formerly classified as 'FRM', will be reclassified as 'agricultural sites' from T_A.

5. Conclusions

In this paper, we addressed the question on vocabulary conflicts in gazetteer integration, using an instance-based thesauri mapping approach to reclassify the loaded objects. We focused specifically on the problem of aligning feature type thesauri.

Our approach used an estimator that creates weighted relationships between terms of distinct thesauri by pre-processing common instances from both gazetteers. To achieve this goal, we collect statistics about the intersection set of instances from gazetteers to be integrated. Then, using the mapping rate estimation model, we assume that all features migrated from one gazetteer to another must to be reclassified using the term from the new feature type thesaurus with the greater mapping rate estimation value, given a threshold (0.4 in our experiments) to be considered an successfully aligned term.

References

ADL (1999), "Alexandria Digital Library Gazetteer", Map and Imagery Lab, Davidson Library, University of California, Santa Barbara. Available at: http://www.alexandria.ucsb.edu/gazetteer

Breitman, K. K., Felicissimo, C. H. and Casanova, M. A. (2005), "CATO – A Lightweight Ontology Alignment Tool". Proc. 17th Conf. on Advanced Information Systems Engineering (CAISE'05), 2005, Porto, Portugal.

Castano, S. et al. (2004), "Semantic Information Interoperability in Open Networked Systems". In: Proc. of the Int. Conference on Semantics of a Networked World (ICSNW), in cooperation with ACM SIGMOD 2004, Paris, France.

Cyc (2005), "OpenCyc Knowledge Base". Available at: http://www.opencyc.org

Doan, A. et al. (2003), "Learning to match ontologies on the Semantic Web". In: The VLDB Journal - The International Journal on Very Large Data Bases, Volume 12, Issue 4, 2003. ISSN: 1066-8888. pp. 303-319.

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). http://www.dlib.org/dlib/january99/hill/01hill.html

ISO-2788 (1986), "Documentation -- Guidelines for the development of monolingual thesauri", International Standard ISO-2788, Second edition -- 1986-11-15.

Janée, G. and Hill, L. L. (2004), "ADL Gazetteer Protocol". Alexandria Digital Library Project. Retrieved Jul 28 2006. Available at http://www.alexandria.ucsb.edu/gazetteer/protocol/

Leme, L.A.P.P. (2006) *Uma arquitetura de software para aplicações de catalogação automática de dados geográficos*. Dissertação de Mestrado. Departamento de Informática, PUC-Rio

McGuinness, D. et al. (2000), "The Chimaera Ontology Environment". In Proceedings of the 17th National Conference on Artificial Intelligence (AAAI), 2000.

Mena, E. et al. (1996), "OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies". In: Proc. of the First IFCIS Int'l Conf. on Cooperative Information Systems, Brussels (Belgium), IEEE, pp. 14-25. Available at: http://sid.cps.unizar.es/PUBLICATIONS/POSTSCRIPTS/coopis96.ps.gz

Mena, E. et al. (2000), "OBSERVER: An Approach for Query Processing in Global Information Systems based on Interoperation across Pre-existing Ontologies", Int'l journal on Distributed And Parallel Databases (DAPD), 8(2):223-272, Kluwer Academic Publishers. Available at: http://sid.cps.unizar.es/PUBLICATIONS/POSTSCRIPTS/dapd00.ps.gz

Noy, N. F. and Musen, M. A. (2003), "The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping". International Journal of Human-Computer Studies, 2003.

Percivall, G. (2003), OpenGIS® Reference Model, Document number OGC 03-040, Version 0.1.3, Open GIS Consortium, Inc.

Spertus, E., Sahami, M. and Buyukkokten, O. (2005), "Evaluating Similarity Measures: A Large-Scale Study in the Orkut Social Network". In: Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, Illinois, USA, August 21-24, 2005. pp.678-684.

UNESCO (1995), "UNESCO Thesaurus". United Nations Educational, Scientific and Cultural Organization, 1995. http://www.ulcc.ac.uk/unesco

WordNet (2005), "WordNet - a lexical database for the English language". Cognitive Science Laboratory, Princeton University, Princeton, NJ – USA. Available at: http://wordnet.princeton.edu