6
Bibliography

ADL, Alexandria digital library gazetteer. Map and Imagery Lab, Davidson Library, Univ. of California, Santa Barbara. Copyright UC Regents, 1999, Last access on Dez 2008 at: http://www.alexandria.ucsb.edu/gazetteer.

S. Bechhofer, F. van Harmelen, J. Hendler, I. Horrocks, D. L. McGuinness, P. F. Patel-Schneider, and L. A. Stein, OWL web ontology language reference, W3C Recommendation, Feb 2004, Last access on Dez 2008 at: http://www.w3.org/TR/owl-ref/.

P. Bernstein and S. Melnik, Model management 2.0: manipulating richer mappings, in *Proc. of the 2007 ACM SIGMOD Int'l. Conf. on Management of Data*, pp. 1–12, 2007.

A. Bilke and F. Naumann, Schema matching using duplicates, in *Proc. of the 21st Int'l. Conf. on Data Engineering*, pp. 69–80, Apr 2005.

D. F. Brauner, M. A. Casanova, and R. L. Milidiú, Mediation as recommendation: an approach to the design of mediators for object catalogs, in *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, volume 4277 of *Lecture Notes in Computer Science*, pp. 46–47, 2006.

D. F. Brauner, M. A. Casanova, and R. L. Milidiú, Towards gazetteer integration through an instance-based thesauri mapping approach, in *Advances in GeoInformatics; VIII Brazilian Symp. on GeoInformatics, GEOINFO*, pp. 235–245, Nov 2007.

D. F. Brauner, C. Intrator, J. C. Freitas, and M. A. Casanova, An instance-based approach for matching export schemas of geographical database Web services, in *Proc. of the IX Brazilian Symp. on GeoInformatics (GEOINFO)*, pp. 109–120, 2007.

D. F. Brauner, A. Gazola, and M. A. Casanova, Adaptative matching of database web services export schemas, in *Proc. of the 10th Int'l. Conf. on Enterprise Information Systems (ICEIS)*, pp. 49–56, 2008.

D. Brickley, R. Guha, and B. McBride, RDF vocabulary description language 1.0: RDF schema, W3C Recommendation, Feb 2004, Last access on Dez 2008 at: http://www.w3.org/TR/rdf-schema.

M. Casanova, K. Breitman, D. Brauner, and A. Marins, Computer, **40**:102, Oct 2007.

J. Chomicki and G. Saake, *Logics for Databases and Information Systems*, chapter 8 - Description Logics for Conceptual data modeling, Springer, 1998.

T. Cover and J. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

H. Do and E. Rahm, COMA: a system for flexible combination of schema

matching approaches, in *Proc. of the 28th Int'l. Conf. on Very Large Data Bases*, pp. 610–621, 2002.

A. Doan, P. Domingos, and A. Y. Halevy, Reconciling schemas of disparate data sources: a machine-learning approach, in *Proc. of the 2001 ACM SIGMOD Int'l. Conf. on Management of Data*, volume 30, pp. 509–520, May 2001.

F. Duchateau, Z. Bellahsène, and E. Hunt, XBenchMatch: a benchmark for XML schema matching tools, in *Proc. of the 33rd Int'l. Conf. on Very Large Data Bases, Demo Sessions: group 1*, pp. 1318–1321, Sep 2007.

H. Eidenberger and C. Breiteneder, Visual similarity measurement with the feature contrast model, in *Proc. of the Storage and Retrieval for Media Databases Conf. (SPIE)*, volume 5021, pp. 64–76, 2002.

H. Eidenberger, Multimedia Systems, **12**:71, 2006.

J. Euzenat and P. Shvaiko, *Ontology matching*, Springer-Verlag, 2007.

W. Frakes and R. Baeza-Yates, *Information retrieval: data structure and algorithms*, Prentice Hall, 1992.

A. Gazola, D. Brauner, and M. A. Casanova, A mediator for heterogeneous gazetteers, in *Poster session of the 22nd Brazilian Symposium on Database*, 2007.

A. Gazola, A software infrastructure for catalog matching, Master's thesis, Departamento de Informática, PUC-Rio, Mar 2008.

GNIS, Geographic names information system, 2005, Last access on Dez 2008 at: http://geonames.usgs.gov.

L. Hill, J. Frew, and Q. Zheng, D-Lib Magazine, **5**, Jan 1999.

D. Hindle, Noun classification from predicate-argument structures, in *Proc. of the 28th Annual Meeting on Association for Computational Linguistics*, pp. 268–275, 1990.

I. Horrocks, P. F. Patel-Schneider, H. Boley, S. Tabet, B. Grosofand, and M. Dean, SWRL: A semantic web rule language combining OWL and RuleML, W3C Member Submission, May 2004, Last access on Dez 2008 at: http://www.w3.org/Submission/SWRL/.

ISO2788, Guidelines for the establishment and development of monolingual thesauri, 1986.

G. Janée and L. L. Hill, The ADL gazetteer protocol, version 1.2, copyright UC Regents, Map and Imagery Lab, Davidson Library, Univ. of California, Santa Barbara, Sep 2004, Last access on Dez 2008 at: http://www.alexandria.ucsb.edu/gazetteer/protocol.

G. Klyne, J. J. Carroll, and B. McBride, Resource description framework (RDF): Concepts and abstract syntax, W3C Recommendation, Feb 2004, Last access on Dez 2008 at: http://www.w3.org/TR/rdf-concepts.

J. Lee, Journal of Documentation, **49**:188, 1993.

L. A. P. Leme, D. F. Brauner, K. K. Breitman, M. A. Casanova, and A. Gazola, Journal Innovations in Systems and Software Engineering, **4**:315, Oct 2008.

L. A. P. P. Leme, M. A. Casanova, K. K. Breitman, and A. L. Furtado, Evaluation of similarity measures and heuristics for simple RDF schema matching, Monografias em Ciência da Computação MCC44/08, Departamento de Informática, PUC-Rio, Oct 2008.

L. A. P. P. Leme, M. A. Casanova, K. K. Breitman, and A. L. Furtado, Database mediation using multi-agent systems, in *Proceedings of the 32nd Annual IEEE Software Engineering Workshop, co-located with the 3rd International Symposium on Leveraging Applications of Formal Methods, Verification and Validation*, Jan 2009.

L. A. P. P. Leme, M. A. Casanova, K. K. Breitman, and A. L. Furtado, Instance-based OWL schema matching, in *Proceedings of the 11th International Conference on Enterprise Information Systems*, May 2009.

D. Lin, An information-theoretic definition of similarity, in *Proc. of the 15th Int'l. Conf. on Machine Learning*, pp. 296–304, 1998.

J. Madhavan, P. A. Bernstein, and E. Rahm, Generic schema matching with Cupid, in *Proc. of the 27th Int'l. Conf. on Very Large Data Bases*, pp. 49–58, 2001.

J. Madhavan, P. Bernstein, A. Doan, and A. Halevy, Corpus-based schema matching, in *Proc. of the 21st Int'l. Conf. on Data Engineering*, pp. 57–68, Apr 2005.

J. Madhavan, S. Cohen, X. L. Dong, A. Y. Halevy, S. R. Jeffery, D. Ko, and C. Yu, Web-scale data integration: You can afford to Pay as You Go, in *Proc. of the 3rd Biennial Conf. on Innovative Data Systems Research (CIDR)*, volume 7, pp. 342–350, 2007.

C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*, The MIT Press, 2002.

S. Melnik, H. Garcia-Molina, and E. Rahm, Similarity flooding: a versatile graph matching algorithm and its application to schema matching, in *Proc. of the 18th Int'l. Conf. on Data Engineering*, pp. 117–128, 2002.

G. Percivall, Open GIS Consortium, Inc., 2003.

E. Prud'hommeaux and A. Seaborne, SPARQL query language for RDF, W3C Recommendation, Jan 2008, Last access on Dez 2008 at: http://www.w3.org/TR/rdf-sparql-query.

W. V. Quine, The Journal of Philosophy, **65**:185, Apr 1968.

E. Rahm and P. Bernstein, The VLDB Journal, **10**:334, 2001.

P. Resnik, Using information content to evaluate semantic similarity in a taxonomy, in *Proc. of the 14th Int'l. Joint Conf. on Artificial Intelligence (AAAI)*, pp. 448–453, 1995.

H. Tang, H. Maitre, and N. Boujemaa, Similarity measures for satellite images with heterogeneous contents, in *Proc. of the Urban Remote Sensing Joint Event*, pp. 1–9, 2007.

A. Tversky and I. Gati, Cognition and categorization, **1**:79, 1978.

O. Udrea, L. Getoor, and R. Miller, Leveraging data and structure in ontology integration, in *Proceedings of the 2007 ACM SIGMOD Intl'l Conf. on Management of data*, pp. 449–460, Jun 2007.

UNESCO, UNESCO Thesaurus, 1995, Last access on Dez 2008 at: http://www.ulcc.ac.uk/unesco.

J. Wang, J. Wen, F. Lochovsky, and W. Ma, Instance-based schema matching for web databases by domain-specific query probing, in *Proc. of the 13th Int'l. Conf. on Very Large Data Bases*, pp. 408–419, Aug 2004.

WordNet, A lexical database for the English language, 2005, Last access on Dez 2008 at: http://wordnet.princeton.edu.

7
Appendix – Setup and calibration of similarity models

The matching approach we described in this thesis represents the elements to be matched with one or more sets of objects. For example, the thesauri terms are represented by sets of instance ids which are classified by the term, numeric properties are represented by sets of observed values of the property and by sets of ordered pairs of the form (instance id, value), character string properties are represented by sets of tokens extracted from their observed values and pairs of the form (instance id, token), and classes are represented by sets of property names. In general, an element *e* to be matched is denoted by a set of sets, i.e. $D_e = \{ S_1, S_2, \ldots, S_n\}$, where we define S_i as a *denotation set*. Two elements *e* and *e'* of the same nature, i.e., two thesauri terms or two numerical properties, must have the same types of denotation sets.

Similarity functions provide means of measuring the similarity between denotation sets. In general, given similarity functions $\sigma_1,...,\sigma_n$, a function $sim:R^n \rightarrow R$ and denotation sets $D_e = \{S_1, S_2, \ldots, S_n\}$ and $D_{e'} = \{S_1', S_2', \ldots, S_n'\}$, the *similarity between e and e'*, expressed as $\Delta(D_e,D_e')$ is defined as the function

$$\Delta(D_e,D_{e'}) = sim(\sigma_1(S_1,S_1'), \sigma_2(S_2,S_2'), \ldots,\sigma_n(S_n,S_n'))$$

The function *sim* may be any function like *max*, *mean*, *weighted mean*, etc, defined on R^n, and σ_i may be any similarity function, such as those presented in Chapter 2. The matching algorithms of Chapter 4 compute the matchings between two elements *e* and *e'* when $\Delta(D_e,D_{e'}) \geq threshold$.

At this point, observe that the similarity $\Delta(D_e,D_{e'})$ depends on the denotation sets D_e and $D_{e'}$ used to represent elements *e* and *e'* (recall that we may consider each denotation set as a multiset or not), the *sim* function, the similarity functions σ_i and the threshold value.

The type of denotation sets adopted, the *sim* function and the similarity functions σ_i are the *similarity model* for the matching algorithms, and the

threshold value is the *calibration* of the similarity model.

The matching approach proposed in Chapter 4 consists of four steps: 1) temporary property matching, 2) class matching, 3) instance matching and 4) refinement of property matching. Each step may use different similarity models and requires a preliminary calibration in order to maximize the performance of the results.

The calibration process requires a training corpus where the matching elements are manually identified and labeled. For each step of the matching approach, the process consists of varying the similarity model and the calibration, and measuring the *overall performance (f)* of the results. Recall that *precision=tp/(tp+fp)*, *recall=tp/(tp+fn)* and *f=2*precision*recall/(precision+recall)*. The best model/calibration for each step is selected.

However, to avoid *overfitting* of the similarity model with respect to the training corpus, we suggest using *cross validation*, which is a process with the following major steps

1. The training corpus is divided in *n* parts.
2. Each similarity model is calibrated with data of *n-1* parts and tested with the remaining *n* part.
3. Steps 1 and 2 are repeated for each of the *n* parts.
4. The final performance of each similarity model is the average overall performance *(f)* for each of the *n* parts.

We describe such an evaluation in (Leme et al. 2008b), where we were interested in evaluating the best similarity model for property matchings. In this experiment, we used data extracted from the gazetteers Alexandria Digital Library and Geonames, and data extracted from eBay and Amazon. The denotation sets for properties were the set of tokens *T* and the set *IV* of pairs of the form *(instance, token)*. We considered both sets as multisets, denoted $\overline{T}$ and $\overline{IV}$, respectively, and as sets in the usual sense, denoted *T* and *IV*. We adopted *max* as the *sim* function, and the *cosine* with *TF/IDF*, the *contrast model function* and the *information theory measure* as similarity functions. For the contrast model function, we used as parameter values α=1.0, and β and $\gamma \in [1.0,10.0]$. The

threshold varied from 0.0 to 1.0, in steps of 0.1.

Table 22 presents the results of the cross validation process. The selected lines indicate the best models. The experiments showed that, for property matching, the similarity function based on the contrast model performs better than the other functions. This result does not represent a final conclusion with respect to similarity models. On the contrary, it should be viewed as a guideline for more elaborate experiments using different similarity models and data.

Table 22. Automatically obtained vocabulary matching from eBay into Amazon

Similarity models					Calibration	Selected model		
σ_i	α	β,γ	sim()	denotation sets	threshold	β,γ	threshold	*f*
contrast model	1.0	[1.0,10.0]	max()	*T,IV*	[0,1.0]	3.5	0.1	60%
contrast model	1.0	[1.0,10.0]	max()	$\overline{T},\overline{IV}$	[0,1.0]	2.5	0.1	53%
contrast model	1.0	[1.0,10.0]	max()	*T*	[0,1.0]	3.5	0.1	65%
contrast model	1.0	[1.0,10.0]	max()	$\overline{T}$	[0,1.0]	2.5	0.1	51%
information theory	-	-	max()	*T,IV*	[0,1.0]	-	0.15	58%
information theory	-	-	max()	$\overline{T},\overline{IV}$	[0,1.0]	-	0.05	52%
information theory	-	-	max()	*T*	[0,1.0]	-	0.1	59%
information theory	-	-	max()	$\overline{T}$	[0,1.0]	-	0.05	50%
cosine with TF/IDF	-	-	max()	*T,IV*	[0,1.0]	-	0.2	56%
cosine with TF/IDF	-	-	max()	$\overline{T},\overline{IV}$	[0,1.0]	-	0.15	57%
cosine with TF/IDF	-	-	max()	*T*	[0,1.0]	-	0.2	56%
cosine with TF/IDF	-	-	max()	$\overline{T}$	[0,1.0]	-	0.15	57%