

Contents

Preface	xi
1 Introduction	1
1.1 Schema Matching in the Data Integration Life Cycle	2
1.2 Manuscript Outline	3
2 Models of Uncertainty	5
2.1 Probability Theory	5
2.2 Fuzzy Set Theory	6
2.2.1 Triangular Norms	6
2.2.2 Fuzzy Aggregate Operators	7
2.3 Discussion	8
3 Modeling Uncertain Schema Matching	9
3.1 Model	9
3.1.1 Schema and attributes	10
3.1.2 Attribute correspondences and the similarity matrix	11
3.1.3 Schema matching	14
3.1.4 (Yet Another) Schema Matcher Classification	15
3.2 Model Usage	19
3.2.1 Deep Web Information	19
3.2.2 Semantic Matching	20
3.2.3 Holistic Matching	20
3.2.4 Non-(1 : 1) cardinality constraints	21
3.3 Reasoning with Uncertain Schema Matching	22
3.3.1 Reasoning using Fuzzy Set Theory	22
3.3.2 Reasoning using Probability Theory	26
3.4 Assessing Matching Quality	27
3.4.1 The monotonicity principle	29

4	Schema Matcher Ensembles	33
4.1	The Art of Matcher Ensembling	33
4.2	2LNB: A Voting Mechanism for Ensembles	36
4.3	Constructing Ensembles	41
4.3.1	AdaBoost	42
4.3.2	SMB: Schema Matcher Boosting	43
4.3.3	Ensemble Construction with SMB: Empirical Analysis	46
5	Top-K Schema Matchings	49
5.1	Top- K Schema Matchings: Definition	49
5.2	Algorithms for Identifying Top- K Matchings	51
5.2.1	Finding the Top-2 Matchings using A_{best}	51
5.2.2	Finding the Top-2 Matchings using Alternating Paths	52
5.2.3	Finding the Top- K Matchings	55
5.3	Extending top- K identification to Ensembles	60
5.3.1	The Matrix-Direct Algorithm	61
5.3.2	Matrix-Direct Algorithm with Bounding	63
5.4	Schema matching verification	65
5.5	Finding probabilistic attribute correspondences	69
6	Applications	71
6.1	Peer-to-Peer Data Integration Systems	71
6.2	Disaster Data Management	72
6.3	Web Service Discovery and Composition	73
6.4	Communities of Knowledge	74
7	Conclusions and Future Work	75
	Author's Biography	85