Schema Matching: The Power of Knowing

Avigdor Gal, Technion – Israel Institute of Technology

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

Uncertain Schema Matching: The Power of not Knowing

Avigdor Gal, Technion – Israel Institute of Technology

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

Peer-to-Peer Data Integration Systems

- The European Future Internet initiative envisions small and medium businesses participating in value chains and networks that emerge dynamically.
- Under this vision, the Internet will meet its promise to become a reliable, seamless and affordable collaboration and sharing platform.
- In principle, the ground is already laid for contemporary enterprises to be able to collaborate flexibly and affordably.
- However, businesses rarely share the same vocabulary and business semantics, raising the costs of B2B interoperability and collaboration.
- Small and medium enterprises cannot afford repeated data integration with new partners for ad-hoc collaboration, leaving the competitive advantage to large enterprises.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Peer-to-Peer Data Integration Systems

NisB: a case in point

- The NisB project (http://www.nisb-project.eu/) aims at easing ad-hoc data integration by harnessing the accumulative connectivity of the Web.
- NisB can help lower the burden for enterprises that deal with multiple heavy-weight and standard-setting enterprises in their value networks.
- It can also help business groups entering a new industrial sector to jointly leverage interoperability efforts.
- Fragments of interoperability information are shared and reused for establishing user-centric understanding of diverse business schemata and vocabularies.
- Bits and pieces of past experiences and practices are re-composed and orchestrated to solve new problems.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Peer-to-Peer Data Integration Systems

Design challenges

- Tolerance towards incomplete, erroneous or evolving information, such as occasional changes in business schemata.
- Quantifiable schema matching uncertainty analysis to allow users to be informed about the usefulness of reusing a matching.
- Keeping context in schema matchings to allow reusability.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

background

- Schema matching is the task of providing correspondences between concepts describing the meaning of data in various heterogeneous, distributed data sources.
- Schema matching research has been going on for more than 25 years now, first as part of schema integration and then as a standalone research field.
- Over the years, a realization has emerged that schema matchers are inherently uncertain.
- A prime reason for the uncertainty of the matching process is the enormous ambiguity and heterogeneity of data description concepts:
  - It is unrealistic to expect a single matcher to identify the correct mapping for any possible concept in a set.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Schema Matching in the Data Integration Life Cycle

- In 2004, Melnik and Bernstein offered a unique contribution to understanding the foundations of schema matching.
  - The *match* operator
  - Schema matching vs. schema mapping

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Overview

Modeling uncertainty

- probability theory [44]
- fuzzy set and fuzzy logic [47]
- lower and upper probabilities [27]
- Dempster-Shafer belief functions [27]
- possibility measures [27]

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Overview

Probability theory vs. fuzzy set theory in schema matching

- Probability theory is a representative of a quantitative approaches in schema matching [12, 20, 8]
- Fuzzy set theory ia a representative of a qualitative approaches in schema matching [19]
- Other approaches:
  - Interval probabilities [34]
  - Probabilistic datalog [43]
  - Information loss estimation [39]

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Probability Theory

Possible world semantics

- An intuitively appealing way to define a probability space [25].
- A probability space is a triple $pred = (W, F, \mu)$ such that:
  - W is a set of possible worlds, with each possible world corresponding to a specific set of event occurrences that is considered possible. A typical assumption is that the real world is one of the possible worlds.
  - $F \subseteq 2^{|W|}$ is a σ-algebra over W. σ-algebra, in general, and in particular F, is a nonempty collection of sets of possible worlds that is closed under complementation and countable unions. These properties of σ-algebra enable the definition of a probability space over F.
  - $\mu : F \rightarrow [0, 1]$ is a probability measure over F.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Fuzzy Set Theory

Definition

- A *fuzzy set* M on a universe set U is a set that specifies for each element $x \in U$ a degree of membership using a membership function

$$\mu_M : U \rightarrow [0, 1]$$

  where $\mu_M(x) = \mu$ is the fuzzy membership degree of the element x in M.
- The degree of membership of an element over all fuzzy sets M does not necessarily sum to 1.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

Detailed example

Sample Database Schema Description

Table: Sample Database Schema Description

Database R				
CardInfo	type	cardNum	...	
HotelInfo	hotelName	neighborhood	...	
Reservations	cardNum	lastName	...	
Database S				
CardInformation	type	cardNum	...	
HotelCardInformation	clientNum	expiryMonth	...	
ReserveDetails	clientNum	name	...	
Database T				
CityInfo	city	neighborhood	...	
Subway	city	station	...	

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Detailed example

Example

- Design of a hotel reservation portal.
- The portal merges various information databases for the hotel chain RoomsRUs
- Mashup application to position hotels on a geographical map.
- Three relational databases:
  - Database R contains three relations: CardInfo, HotelInfo, and Reservations.
  - Database S also stores credit card information, distinguishing between hotel credit cards and major credit cards.
  - Database T has two relations: CityInfo and Subway.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Schema and attributes

- *schema* $S = \{A_1, A_2, ..., A_n\}$ is a finite set of *attributes*.
- Attributes can be both simple and compound.
- Examples:
  - Simple attributes: lastName, firstName.
  - Compund attributes: creditCardInfo, combining type, cardNum, and expiry
- Is this representation too simple?
  - Metadata models use complex representations: relational databases use tables and foreign keys, XML structures have hierarchies, OWL ontologies contain general axioms.
  - Modeling the uncertainty in matching attributes, no richer representation of data models is needed.
  - Example 1: matching lastName and firstName does not require representing their composition in a compound attribute called name.
  - Example 2: when matching compound structures such as XML paths, XML paths **are** the elements we define as attributes in our schemata.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Attribute correspondences and the similarity matrix

- Let S and S' be schemata with n and n' attributes, respectively.
- Let $\mathcal{S} = S \times S'$ be the set of all possible *attribute correspondences* between S and S'.
  - $\mathcal{S}$ is a set of attribute pairs.
  - Example: (arrivalDate, checkInDay)).
- Let $M\,(S, S')$ be an $n \times n'$ *similarity matrix* over $\mathcal{S}$.
  - $M_{i,j}$ represents a degree of similarity between the i-th attribute of S and the j-th attribute of S'.
  - $M_{i,j}$ is a real number in $(0, 1)$.
- $M\,(S, S')$ is a *binary similarity matrix* if for all $1 \leq i \leq n$ and $1 \leq j \leq n'$, $M_{i,j} \in \{0, 1\}$.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Attribute correspondences and the similarity matrix

Table: A Similarity Matrix Example

$S_1 \longrightarrow$ $\downarrow S_2$	1 cardNum	2 city	3 arrivalDate	4 departureDate
1 clientNum	0.843	0.323	0.317	0.302
2 city	0.290	1.000	0.326	0.303
3 checkInDay	0.344	0.328	0.351	0.352
4 checkOutDay	0.312	0.310	0.359	0.356

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Attribute correspondences and the similarity matrix

Table: A Binary Similarity Matrix Example

$S_1 \longrightarrow$ $\downarrow S_2$	1 cardNum	2 city	3 arrivalDate	4 departureDate
1 clientNum	1	0	0	0
2 city	0	1	0	0
3 checkInDay	0	0	0	1
4 checkOutDay	0	0	1	0

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Schema matchers

- Similarity matrices are generated by schema matchers.
- *Schema matchers* are instantiations of the schema matching process.
- They differ mainly in the measures of similarity they employ, which yield different similarity matrices.
  - arbitrarily complex
  - use various techniques.
    - similar attributes are more likely to have similar names [29, 46].
    - similar attributes share similar domains [33, 21].
    - instance similarity as an indication of attribute similarity [5, 10].
    - experience of previous matchings as indicators of attribute similarity [28, 32, 46].

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Matcher examples

Example

- Term matching compares attribute names to identify syntactically similar attributes.
- To achieve better performance, names are preprocessed using several techniques originating in IR research.
- Term matching is based on either complete words or string comparison.
- Example: relations CardInfo and HotelCardInformation.
  - The maximum common substring is CardInfo
  - the similarity of the two terms is $\frac{length(\text{CardInfo})}{length(\text{HotelCardInformation})} = \frac{8}{20} = 40\%$.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Matcher examples

Example

- Value matching utilizes domain constraints (*e.g.*, drop lists, check boxes, and radio buttons).
- It becomes valuable when comparing two attributes whose names do not match exactly.
- Example, attributes arrivalDate and checkInDay.
  - Associated value sets $\{(Select),1,2,...,31\}$ and $\{(Day),1,2,...,31\}$, respectively.
  - content-based similarity is $\frac{31}{33} = 94\%$
  - Significantly higher than their term similarity ($\frac{2(\text{Da})}{11(\text{arrivalDate})} = 18\%$).

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Matcher examples

Example

- A composite attribute is composed of other attributes (either atomic or composite).
- Composition can be translated into a hierarchy.
- Composition matcher assigns similarity to attributes based on the similarity of their neighbors.
- The Cupid matcher [33], for example, is based on attribute composition.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Matcher examples

Example

- The order in which data are provided in an interactive process is important.
  - Data given at an earlier stage may restrict the options for a later entry.
  - Example: filling in a form on a hotel reservation site.
- Precedence relationships can be translated into a precedence graph.
- Precedence matcher is based on *graph pivoting*.
- When matching two attributes, each is considered to be a pivot within its own schema.
- By comparing preceding subgraphs and succeeding subgraphs, the confidence strength of the pivot attributes is determined.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Similarity Matrices Explained

- It was hypothesized and empirically validated by [35] that when encoding attribute pair similarities in a similarity matrix, a matcher would be inclined to assign a value of 0 to each pair it conceives not to match, and a similarity measure higher than 0 (and probably closer to 1) to those attribute pairs that are conceived to be correct.
- This tendency, however, is masked by "noise," whose sources are rooted in missing and uncertain information.
- instead of expecting a binary similarity matrix the values in a similarity matrix form two probability distributions over $[0, 1]$.

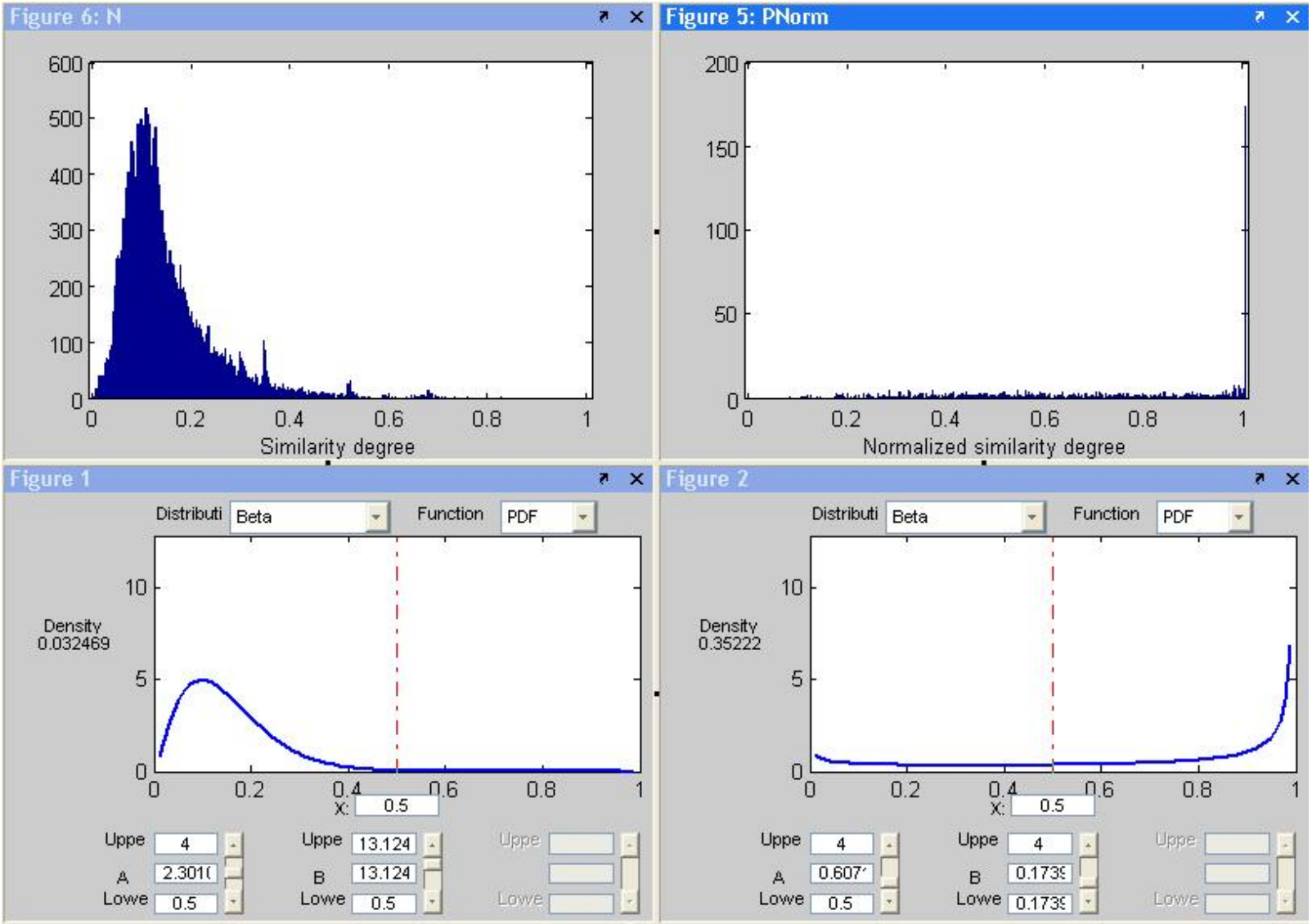
Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Similarity Matrices Explained

- Evaluation of more than 26,000 attribute pairs coming from 50 schema pairs.
- Attribute pairs separated to pairs reflecting attribute correspondences and those that not.
- The precedence matcher was used.
- Similarity values of correct correspondences normalized.
- Beta distribution estimations were added:
  - The beta distribution can be used to model a random phenomenon whose set of possible values is in some finite interval $[c, d]$.
  - A beta distribution has two tuning parameters, a and b.
  - $b > a$ for left-skewed density function.
  - $a > b$ for right-skewed density functions.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Similarity Matrices Explained



Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Similarity Matrices

Schema Matching

Schema Matching Classification

Model Usage

Assessing Matching Quality

Schema Matcher Ensembles

Top-K

Similarity Matrices: Discussion

- Schema matchers often use data model semantics when determining the similarity between attributes.
  - XML structure has been used in Cupid [33] to support or dispute linguistic similarities.
  - Similarity flooding [38] uses structural links between attributes to update linguistic similarities.
- Once this similarity has been determined and recorded in the similarity matrix, the original semantics that derived it is no longer needed.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Schema Matching: Preliminaries

- $\Sigma = 2^{\mathcal{S}}$ is the power-set of all possible *schema matchings* between the schema pair (S, S'), where a schema matching $\sigma \in \Sigma$ is a set of attribute correspondences.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Similarity Matrices

Schema Matching

Schema Matching Classification

Model Usage

Assessing Matching Quality

Schema Matcher Ensembles

Top-K

Schema Level Constraints

- $\Gamma : \Sigma \rightarrow \{0, 1\}$ is a boolean function that captures the application-specific constraints on schema matchings:
  - cardinality constraints.
  - inter-attribute correspondence constraints
- Γ partitions Σ into two sets, valid and invalid:
  - The set of all *valid* schema matchings in Σ is given by $\Sigma_\Gamma = \{\sigma \in \Sigma \mid \Gamma(\sigma) = 1\}$
  - (go to Slide 2.5)
- Γ is a general constraint model, where $\Gamma(\sigma) = 1$ means that the matching σ can be accepted by a designer.
- Γ has been modeled in the literature using special types of matchers called *constraint enforcers* [31], whose output is recorded in a binary similarity matrix.
- Γ is a *null constraint function* if for all $\sigma \in \Sigma$, $\Gamma(\sigma) = 1$.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Schema Matching Process: Functional Definition

- Input: two schemata S and S' and a constraint boolean function Γ.
- Output: a *schema matching* $\sigma \in \Sigma_\Gamma$.

Example

- Table 2 represents a step in the schema matching process, in which the similarity of attribute correspondences is recorded in a similarity matrix.
- The similarity matrix in Table 3 presents a possible outcome of the matching process, where all attribute correspondences for which a value of 1 is assigned are part of the resulting schema matching.
- The constraint function that is applied in this example enforces a $1:1$ matching.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Schema Matching and Matrix Satisfaction

Definition (Matrix Satisfaction)

Let $M(S, S')$ be an $n \times n'$ similarity matrix over $\mathcal{S}$. A schema matching $\sigma \in \Sigma$ is said to *satisfy* $M(S, S')$ (denoted $\sigma \models M(S, S')$) if $\left(A_i, A'_j\right) \in \sigma \rightarrow M_{i,j} > 0$.
$\sigma \in \Sigma_\Gamma$ is said to *maximally satisfy* $M(S, S')$ if $\sigma \models M(S, S')$ and for each $\sigma' \in \Sigma_\Gamma$ such that $\sigma' \models M(S, S')$, $\sigma' \subset \sigma$.

- The output of a schema matching process is $M(S, S')$.
- An attribute pair (A_i, A_j) is an attribute correspondence in the output schema matching only if $M(i', j) > 0$.
- A schema matching σ satisfies M if the above is true for any attribute pair in σ.
- the output of the schema matching process is a valid schema matching that **maximally** satisfies $M(S, S')$.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Matrix Satisfaction: Discussion

- Σ is partitioned into two sets, based on satisfaction.
- Partition is based on the application and the matcher's ability to determine attribute requirements.
- compare with Γ partitioning of Σ in Slide 2.2, the two may not overlap.
- Maximal satisfaction is defined over valid schema matchings only:
  - In the absence of a constraint function Γ, *i.e.*, whenever $\Sigma_\Gamma = \Sigma$ or whenever Γ is ignored by the matcher, there is exactly one schema matching that maximally satisfies M.
    - This schema matching contains all attribute correspondences (A_i, A'_j) for which $M_{i,j} > 0$.
  - When Γ is both meaningful and used by the matcher, there may not be exactly one valid schema matching that maximally satisfies M.
    - Example: a $1:1$ constraint.
    - Example: none of the valid schema matchings satisfy M.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Second-Line Matchers

Definition

Given schemata S and S', we denote by $\mathcal{M}(S, S')$ the (possibly infinite) set of similarity matrices $M(S, S')$. A second-line schema matcher

$$SM : \mathcal{M}(S, S')^{+} \times \Gamma \rightarrow \mathcal{M}(S, S')$$

is a mapping, transforming one (or more) similarity matrices into another similarity matrix.

- A second-line matcher (2LM) is a schema matcher whose inputs are no longer the schemata S and S', but rather a similarity matrix $M(S, S')$ (together with Γ).
- *First-line schema matchers* (1LM) operate on the schemata themselves, using semantics of the application.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Second-Line Matchers: examples

Example

- Combined is a weighted combination of the Term, Value, Composition, and Precedence matchers.
- The Maximum Weighted Bipartite Graph (MWBG) algorithm and the Stable Marriage (SM) algorithm both enforce a cardinality constraint of $1:1$.
  - MWBG uses a bipartite graph, where nodes in each side of the graph represent attributes of one of the schemata, and the weighted edges represent the similarity measures between attributes. Algorithms such as those presented in [23] provide the output of the MWBG heuristic.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Second-Line Matchers: examples

Example

- (MWBG) and SM (cont.):
  - SM takes a similarity matrix and applies a stable marriage algorithm [26] to identify a schema matching.
  - Intersection [36] computes and outputs the intersection set of both algorithm outputs.
  - Union includes in the output matching any attribute correspondence that is in the output of either MWBG or SM.
  - Intersection and Union do not enforce $1:1$ matching.
- Dominants chooses *dominant pairs*, those pairs in the similarity matrix with maximum value in both their row and their column.
- 2LNB [35] is a 2LM that uses a naïve Bayes classifier over matrices to determine attribute corespondences.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Second-Line Matchers: examples

Example (eTuner)

- A model of a $1:1$ matching system was defined by [31] to be a triple, one element being a library of matching components.
- This library has four types of components: Matcher, Combiner, Constraint Enforcer, and Match Selector.
  - *Matcher* is a 1LM, in its classical definition.
  - *Combiner* [9] follows the definition of a schema matcher with a null constraint function.
  - *Constraint enforcer* is a 2LM (current definition in Section 2 allows adding constraints at 1LM as well)
  - *Match selector* returns a matrix in which all elements that are not selected are reduced to 0, *e.g.* using thresholding and MWBG.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Second-Line Matchers: examples

Example (Top-K)

- A heuristic that utilizes the top-K best schema matchings to produce an improved schema matching [18].
- It is a special type of a combiner and a match selector.
- The input does not come from different matchers.
- Multiple matrices by same schema matcher are aggregated to a single similarity matrix by thresholding.

Discussion

- The modeling of 2LM can serve as a reference framework for comparing various research efforts in schema matching.
- Example: Combiners and match selectors were combined and redefined by [18].
- 2LMs aim at improving the outcomes of 1LMs.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Schema Matcher Classification

Table: Two dimension matcher classification

Matcher	**1LM**	**2LM**
Non-decisive	Term	Combined
Decision maker		MWBG

- The first dimension separates first- from second-line schema matchers.
- The second dimension separates those matchers that aim at specifying schema matchings (*decision makers*) from those that compute similarity values yet do not make decisions at the schema level.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Schema Matcher Classification

- A matcher is decisive if it satisfies Γ.
- Most common: non-decisive first-line matcher.
- Combiners (COMA's term) are non-decisive 2LMs.
  - Combine similarity matrices of other matchers
  - The similarity matrix is not meant to be used to decide on a single schema matching.
- Decisive 2LMs: algorithms like MWBG and SM.
  - Both are constraint enforcers [31].
  - Both enforce a cardinality constraint of $1:1$.
- First-line decision makers contains few if any matchers.
  - "The comparison activity focuses on primitive objects first...; then it deals with those modeling constructs that represent associations among primitive objects." [3]
  - This dichotomy has in the main been preserved in schema matching as well.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Deep Web Information

- The deep Web hides information behind Web forms.
- This information is stored in databases on the server side.
- The database schemata are not known but the information they contain is exposed to users.
- This information can be extracted and matched.
- matrix rows represent the fields of one form and matrix columns those of the other.
- Each entry in the matrix represents the similarity of two fields.
- Possible matchers:
  - Linguistic similarity of labels
  - Linguistic similarity of field names
  - Domain similarity.
  - structural similarities: composition and precedence.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Semantic Matching

- Pairwise matching results in an ontological relationship (*e.g.*, equivalence and subsumption) between attributes.
- The S-Match system [24] takes two graph-like structures (*e.g.*, XML schemata) as input.
- S-Match generates a confidence measure for semantic relationship being equivalence, subsumption, *etc*.
- For each pair of elements only one relationship holds.
- Matrix representation:
  - Each ontological relationship in a separate matrix..
  - A second-line matcher (specific to S-Match) generates a set of binary matrices, using some thresholds.
  - A constraint enforcer uses a lattice structure (where, for example, equivalence is higher than subsumption) to determines which values are to remain 1.
  - An entry for which a 0 value is recorded in all matrices is assigned 1 for the *idk* (stands for "I don't know") matrix.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

Holistic Matching

- Rather than matching two schemata at a time, holistic matching matches multiple schemata in an attempt to reach a consensus terminology for a domain.
- Representation: multi-dimensional matrices.
- Example: holistic matching of three schemata uses a 3-dimensional matrix where each entry represents the certainty of three attributes, one from each schema, being matched.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Similarity Matrices
Schema Matching
Schema Matching Classification
Model Usage
Assessing Matching Quality
Schema Matcher Ensembles
Top-K

What is the question?

- What qualifies a schema matcher to be considered "good"?
- Empirical, explanatory analysis, testing schema matchers using *a posteriori* metrics.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Precision and Recall

- The evaluation of schema matchings is performed with respect to an *exact matching*, based on expert opinions.
- *Precision* and *recall* are used for the empirical evaluation of performance.
- Given $n \times n'$ attribute pairs:
  - $c \leq n \times n'$ attribute correspondences, with respect to the exact matching.
  - $t \leq c$ is the number of pairs, out of the correct correspondences, that were chosen by a matcher.
  - $f \leq n \times n' - c$ is the number of incorrect correspondences.
- Precision:
$$P = \frac{t}{t+f}$$
- Recall:
$$R = \frac{t}{c}$$
- Higher values of both precision and recall are desired.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Precision and Recall

- Derivatives of precision and recall:
  - *F-Measure*:
    - The harmonic mean of precision and recall.
    - Formula:
$$FM = 2 \cdot \frac{PR}{P + R}$$
  - *overall*:
    - Evaluates post-match effort, including the amount of work needed to add undiscovered matchings and remove incorrect matchings.
    - Formula:
$$OV = R \cdot \left(2 - \frac{1}{P}\right) = \frac{t - f}{c}$$
    - This measure may be assigned with negative values.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Precision and Recall

- Derivatives of precision and recall (cont.):
  - *error*:
    - It was used by [41] for schema matching, following [15].
    - Formula:
$$ER = 1 - \frac{(1 + b^2)PR}{b^2 P + R}$$
    - b is a tunable parameter.
    - The lower the value of ER, the better the match.
  - *information loss:* [39]
    - Quantifies the uncertainty that arises in the face of possible semantic changes when translating a query across different ontologies.
    - Formula:
$$LS = 1 - \left(\frac{1}{\gamma P^{-1} + (1 - \gamma) R^{-1}} \right)$$
    - γ is a tunable parameter.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Precision and Recall

- Relationships between measures:
  $FM = 1 - LS$ for $\gamma = 0.5$
  $ER = LS$ for $\gamma = 1$ and $b = 0$
  $ER = 1 - FM$ for $b = 1$
- Precision, recall, and their derivatives have traditionally served the research community to empirically test the performance of schema matchers.
- These metrics are explanatory in nature, measuring the goodness-of-fit of the heuristic to the data.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Soundness and completeness

- Precision and recall provide a form of pragmatic (*a posteriori*) soundness and completeness. Therefore, an exact matching is needed to measure
- semantic soundness and completeness of Schema matchings [4] using a complete ontology.
- lossless mapping and information capacity [30, 40, 2, 7] measure the ability to reconstruct the original data.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

The monotonicity principle

- The monotonicity measure [19], provides a relationship between the behavior of a given matcher and its true performance.
- Instead of just observing the final outcome provided by the matcher, the monotonicity principle observes the internal mechanism that leads to a matcher's decision.
- In that sense, it offers a deeper understanding of a matcher's capability.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Matcher monotonicity definition

- Equivalence schema matching classes on $2^{\mathcal{S}}$:
- Two matchings σ' and σ'' belong to a class p if $P(\sigma') = P(\sigma'') = p$, where $p \in [0, 1]$.
- For each two matchings σ' and σ'', such that $P(\sigma') < P(\sigma'')$, we can compute their schema matching level of similarity, $\Omega(\sigma')$ and $\Omega(\sigma'')$.

Definition

A matching algorithm is *monotonic* if for any two matchings $\{\sigma', \sigma''\} \subseteq 2^{\mathcal{S}}$, $P(\sigma') < P(\sigma'') \rightarrow \Omega(\sigma') < \Omega(\sigma'')$.

- Intuitively, a matching algorithm is monotonic if it ranks all possible schema matchings according to their level of precision.

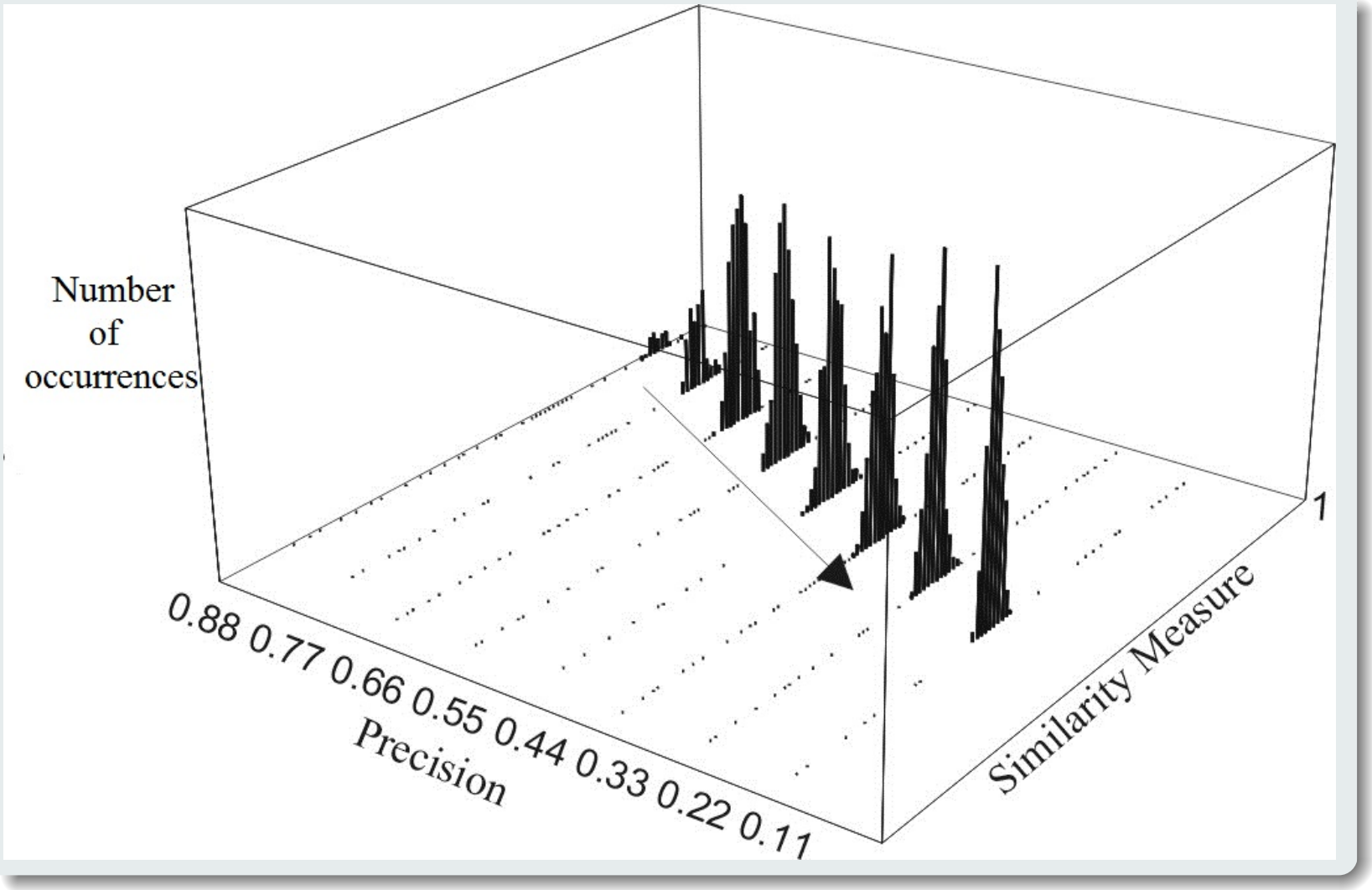
Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Matcher monotonicity explained

- A matching algorithm is monotonic if it ranks all possible schema matchings according to their level of precision.
- A monotonic matching algorithm easily identifies the exact matching:
  - Let σ^* be the exact matching, then $P(\sigma^*) = 1$.
  - For any other matching σ', $P(\sigma') < P(\sigma^*)$.
  - Therefore, if $P(\sigma') < P(\sigma^*)$ then from monotonicity $\Omega(\sigma') < \Omega(\sigma^*)$.
  - All one has to do then is to devise a method for finding a matching σ^* that maximizes Ω.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Matcher monotonicity illustrated



Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Matcher monotonicity explained

- The figure provides an illustration of the monotonicity principle using a matching of a simplified version of two Web forms.
- Both schemata have nine attributes, all of which are matched under the exact matching.
- Given a set of matchings, each value on the x-axis represents a class of schema matchings with a different precision.
- The z-axis represents the similarity measure.
- The y-axis stands for the number of schema matchings from a given precision class and with a given similarity measure.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Matcher monotonicity: insights

- The similarity measures of matchings within each schema matching class form a "bell" shape, centered around a specific similarity measure.
- This behavior indicates a certain level of robustness, where the schema matcher assigns similar similarity measures to matchings within each class.

- The "tails" of the bell shapes of different classes overlap.
- Therefore, a schema matching from a class with lower precision may receive a higher similarity measure than one from a class with higher precision.
- Contradiction to monotonicity.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Statistical monotonicity

Definition (Statistical monotonicity)

Let $\Sigma = \{\sigma_1, \sigma_2, ..., \sigma_m\}$ be a set of matchings over schemata S_1 and S_2 with n_1 and n_2 attributes, respectively, and define $n = \max(n_1, n_2)$. Let $\Sigma_1, \Sigma_2, ..., \Sigma_{n+1}$ be subsets of Σ such that for all $1 \leq i \leq n+1$, $\sigma \in \Sigma_i$ iff $\frac{i-1}{n} \leq P(\sigma) < \frac{i}{n}$. We define M_i to be a random variable, representing the similarity measure of a randomly chosen matching from Σ_i. Σ is *statistically monotonic* if the following inequality holds for any $1 \leq i < j \leq n+1$:

$$\bar{\Omega}(M_i) < \bar{\Omega}(M_j)$$

where $\bar{\Omega}(M)$ stands for the expected value of M.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Statistical monotonicity

- A schema matching algorithm is statistically monotonic with respect to two given schemata if the expected certainty increases with precision.
- Statistical monotonicity can help explain certain phenomena in schema matching:
  - It can explain the lack of "industrial strength" [6] schema matchers.
  - It serves as a guideline as we seek better ways to use schema matchers.
  - It helps understanding why schema matcher ensembles work well
  - It serves as a motivation for seeking top-K matchings.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
The monotonicity principle
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Motivation

- In an effort to increase the robustness of individual matchers in the face of matching uncertainty, researchers have turned to schema matcher ensembles.
- Ensembles combine different schema matchers that use complementary principles to judge the similarity between concepts.
- An ensemble of complementary matchers can potentially compensate for the weaknesses of any given matcher in the ensemble.
- Several studies report on encouraging results when using schema matcher ensembles (*e.g.*, [9, 14, 21, 33, 42]).
- Tools developed for ensemble design include eTuner [31], LSD [10] and OntoBuilder [35, 37].

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The Art of Matcher Ensembling

- A *schema matching ensemble* is a set of schema matchers.
- An ensemble aggregates the similarities assigned by individual matchers to reason about the resulting aggregated ranking of alternative matchings.
- Such an aggregation can be modeled in various ways:
  - A cube, aggregated into a matrix by aggregating the similarity values of each correspondence across ensemble members. [9]
  - Extended by analyzing the relationships between local and global aggregators. [11]
    - Local aggregators combine the similarity measures of attribute correspondences into a schema matching similarity measure by a single matcher.
    - Global aggregator combine the similarity measures of multiple matchers.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The Art of Matcher Ensembling

Ensemble Design Dimensions

Table: Ensemble design dimensions

Participation → / ↓ Execution	Single	Multiple
Sequential		[13]
Parallel	[18]	[9]

Participation dimension

- Works in the literature typically construct matcher ensembles from multiple matchers.
- top-k matching combines the input of a single matcher with different settings.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The Art of Matcher Ensembling

Execution dimension

- To date, *parallel ensembling* dominantes the ensemble research:
  - Combine the judgments of multiple matchers (a similarity cube) into a single matcher (a similarity matrix).
  - eTuner [31] tunes the weights of the different matchers, giving greater weight to more effective matchers.
- Less common is the *sequential ensemble approach*:
  - Matchers are added to an ensemble sequentially, based on the outcomes of earlier stages.
  - Allows matchers to suggest correspondences in "regions" of the similarity matrix in which they "feel" more confident.
  - A matcher can identify correspondences for which it is less confident and pass them on to another matcher.
  - A matcher needs to be able to identify its "strong regions."
  - An example of this line of work is that of [13], which introduces a decision tree to combine matchers.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The Art of Matcher Ensembling

Aggregation dimension

- Linear aggregation.
- Non-linear aggregation, working directly with a global aggregator [1].

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

Impact of Matcher Ensembling: an Empirical Example

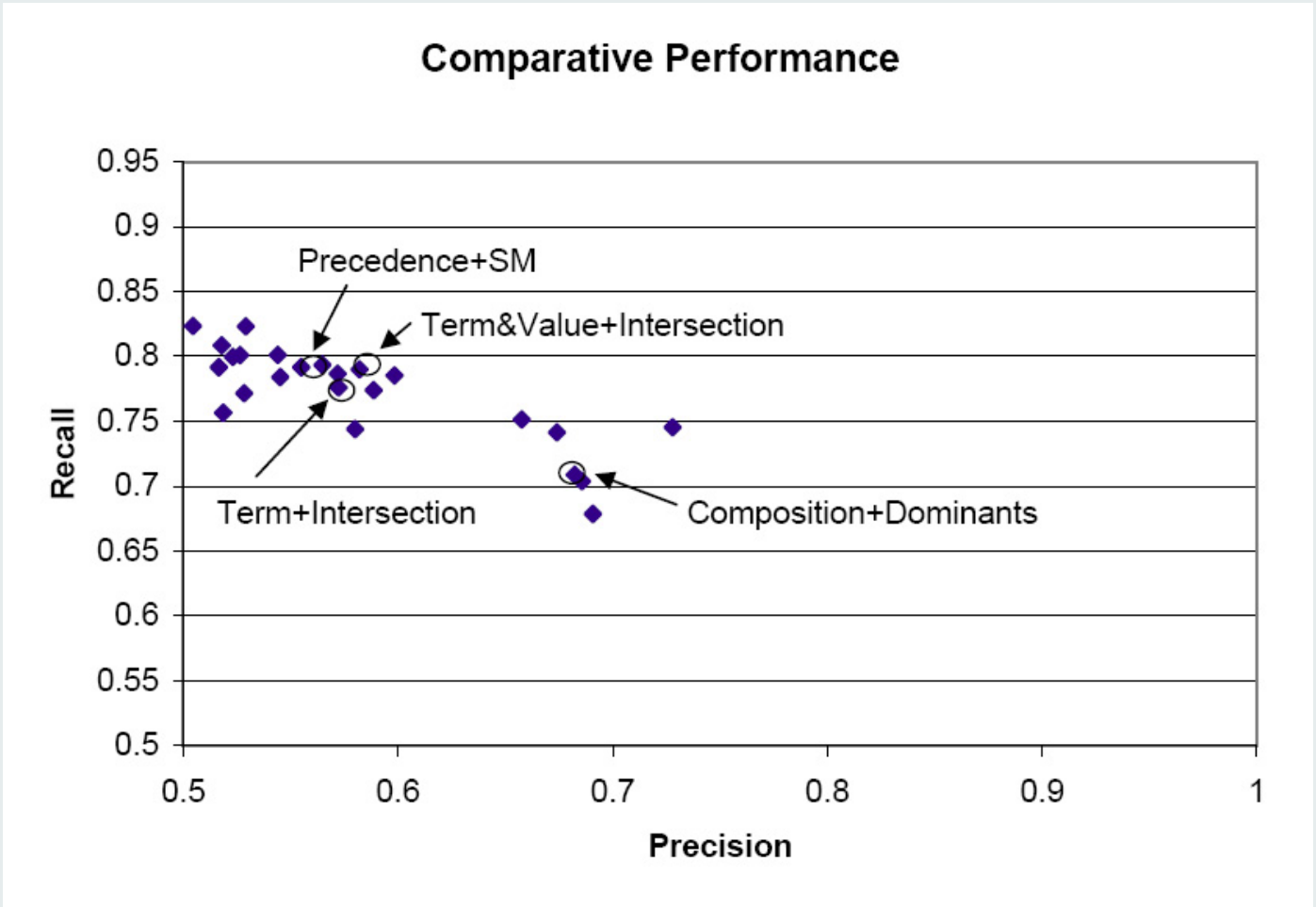


Figure: Relative matcher weights in SMB and individual performance

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

Constructing Ensembles

- Choosing among schema matchers is far from trivial.
  - The number of schema matchers is continuously growing, and this diversity by itself complicates the choice of the most appropriate tool for a given application domain.
  - Empirical analysis shows that there is not (and may never be) a single dominant schema matcher that performs best, regardless of the data model and application domain [19].
- Most research work devoted to constructing ensembles deals with setting the relative impact of each participating matcher. For example, Meta-Learner [10] aims at a weighted average of the decisions taken by the matchers in an ensemble, using least-square linear regression analysis [10]

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

Boosting Ensembles

- Research has shown that many schema matchers perform better than random choice.
- We argue that any (statistically) monotonic matcher is a *weak classifier* [45]—a classifier that is only slightly correlated with the true classification.
- A *weak classifier* for binary classification problems is any algorithm that achieves a weighted empirical error on the training set which is bounded from above by $1/2 - \gamma, \gamma > 0$ for some distribution on the dataset (the dataset consists of weighted examples that sum to unity).
- A weak classifier can produce a hypothesis that performs at least slightly better than random choice.
- The theory of weak classifiers has led to the introduction of *boosting* algorithms (*e.g.*, [45]) that can strengthen weak classifiers to achieve arbitrarily high accuracy.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

AdaBoost

- The AdaBoost algorithm [17] is the most popular and historically most significant boosting algorithm.

Algorithm 1 Boosting

Input: $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, and a space hypotheses $\mathcal{H}$.
/* $\forall 1 \leq i \leq m, x_i \epsilon \mathcal{X}$, and $\forall 1 \leq i \leq m, y_i \epsilon \{-1, +1\}$ */
/* initialization: */
for all $1 \leq i \leq m$ **do**
  $D_1(i) = 1/m$
end for
$t = 1$
repeat
  /* training phase: */
  Find the classifier $h_t : \mathcal{X} \rightarrow \{-1, +1\}$, $h_t \epsilon \mathcal{H}$ that minimizes the error with respect to the distribution D_t: $h_t = \arg_{h_j} \min \varepsilon_j$.
  if $\varepsilon_t \leq 0.5$ **then**
    Choose $\alpha_t \epsilon R$. $\alpha_t = \frac{1}{2} \ln \frac{1-\varepsilon_t}{\varepsilon_t}$
    Update $D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ where Z_t is a normalization factor
    $t = t + 1$
  end if
until $t = T$ or $\varepsilon_t > 0.5$
/* upon arrival of a new instance: */
Output the final classifier: $H(x) = sign(\sum_{k=1}^{\min(t,T)} \alpha_k h_k(x))$

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

AdaBoost explained

- The input to a boosting algorithm is a **set** of m examples where each example (x_i, y_i) is a pair of an instance x_i and the classification of the instance mapping, y_i.
- y_i typically (though not always) accepts a binary value in $\{-1, +1\}$, where -1 stands for an incorrect classification and $+1$ stands for a correct classification.
- The last input element is a hypothesis space $\mathcal{H}$, a set of weak classifiers.
- The algorithm works iteratively. In each iteration the input set is examined by all weak classifiers.
- From iteration to iteration the relative weight of examples changes.
- Most weight is set on the examples most often misclassified in preceding steps.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

AdaBoost explained (cont.)

- The algorithm assign an initial equal weight to all examples. Weights are updated later
- Weak classifiers are applied in parallel, looking for the most accurate h_t over the weighted examples.
- The amount of error of each weak classifier is computed.
- The error measure is in general proportional to the probability of incorrectly classifying an example under the current weight distribution ($\Pr_{i \sim D_t}(h_t(x_i) \neq y_i)$).
- At round t, the weak classifier that minimizes the error measure of the current round is chosen.
- The stop condition limits the amount of error to no more than 50%.
- The stop condition also restricts the maximum number of iterations.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

AdaBoost explained (cont.)

- The amount of change to example weights α_t is determined in a way that reduces error most rapidly (in a greedy way) [16] by minimizing

$$Z_t = \sum_{i=1}^{m} D_t(i) e^{-\alpha_t y_i h_t(x_i)}.$$

- The learned weights are then used to classify a new instance x, by producing $H(x)$ as a weighted majority vote, where α_k is the weight of the classifier chosen in step k and $h_k(x)$ is the decision of the classifier of step k.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

from AdaBoost to ensemble tuning

- The boosting algorithm is merely a shell, serving as a framework for many possible instantiations.
- What separates a successful instantiation from a poor one is the selection of three elements:
  - the instances (x_i)
  - the hypothesis space $(\mathcal{H})$
  - the error measure (ε_t).

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The example set

- The example set $\{(x_i, y_i)\}$ consists of a set of attribute pairs (x_i is a pair!), comprising one attribute from each schema and belonging to the classification of the instance mapping y_i.
- Such a pair represents an attribute correspondence.
- Each instance x_i can be correct (*i.e.*, belonging to the exact matching) or incorrect.
- Therefore, y_i can have two possible values: $(+1)$ (for a correct matching) and (-1) (for an incorrect matching).
- This approach can be easily extended to select multiple attributes from each schema, as long as the matcher itself can assess the similarity measure of multiple attributes.
- For holistic matching, examples can be designed to be sets of attributes from multiple schemata rather than a pair.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The hypothesis space

- Choosing the hypothesis space is more tricky.
- The input to SMB is a similarity matrix $M(S, S')$ (together with Γ, the constraint enforcer function).
- Given schemata S and S', we denote by $\mathcal{M}(S, S')$ the (possibly infinite) set of similarity matrices $M(S, S')$.
- The SMB heuristic is a mapping

$$\text{SMB} : \mathcal{M}(S, S')^* \times \Gamma \to \mathcal{M}(S, S'),$$

  transforming one (or more) similarity matrices into another similarity matrix.
- Elements of the hypothesis space are matrices.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The hypothesis space: cont.

- The most promising hypothesis space seems to be a set of second-line matchers of the type decision makers (whose output is a binary matrix).
- For example, a hypothesis h in $\mathcal{H}$ is (Term, Dominants), where the Dominants second-line matcher is applied to the outcome of the Term first-line heuristic.
- SMB is also a decision maker and the outcome of SMB is a binary matrix.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The error measure ε

- A matcher can either determine a correct attribute matching to be incorrect (false negative) or it can determine an incorrect attribute matching to be correct (false positive).
- Let A_t denote the total weight of the false negative examples, C_t the total weight of the false positive examples, and B_t the total weight of the true positive examples, all in round t.
- Typically, one would measure error in schema matching in terms of precision and recall, translated into boosting terminology as follows:

$$P(t) = \frac{B_t}{C_t + B_t}; R(t) = \frac{B_t}{A_t + B_t} \quad (1)$$

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

The error measure ε: cont.

- These are combined using F-Measure:

$$FM(t) = \frac{2B_t}{A_t + C_t + 2B_t} \tag{2}$$

- A plausible error measure for the SMB heuristic is:

$$\varepsilon_t = 1 - FM(t) = 1 - \frac{2B_t}{A_t + C_t + 2B_t} = \frac{A_t + C_t}{A_t + C_t + 2B_t} \tag{3}$$

- Empirical evaluation suggests that Eq. 3 performs better than other error measures.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

Example of SMB

Example

This example is due to [22].

- Given the hypothesis space $\mathcal{H}$ as described above, and given a dataset of size 70, the SMB heuristic performs 5 iterations.
- First, it creates a dataset with equal weight for each mapping.
- In the first iteration, it picks (Composition, Dominants), which yields the most accurate hypothesis over the initial weight distribution ($\varepsilon_1 = 0.328 \Rightarrow \alpha_1 = 0.359$).
- In the second iteration, the selected hypothesis is (Precedence, Intersection) with $\varepsilon_2 = 0.411$ and $\alpha_2 = 0.180$.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

Example of SMB: cont.

Example

- In the third, (Precedence, MWBG) is chosen with $\varepsilon_3 = 0.42 \Rightarrow \alpha_3 = 0.161$.
- The fourth hypothesis selected (Term and Value, Intersection), with $\varepsilon_4 = 0.46$ and $\alpha_4 = 0.080$.
- The fifth and final selection is (Term and Value, MWBG), with $\varepsilon_5 = 0.49 \Rightarrow \alpha_5 = 0.020$.
- In the sixth iteration no hypothesis performs better than 50% error, so the training phase is terminated after 5 iterations, each with strength α_t.
- The outcome classification rule is a linear combination of the five weak matchers with their strengths as coefficients.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

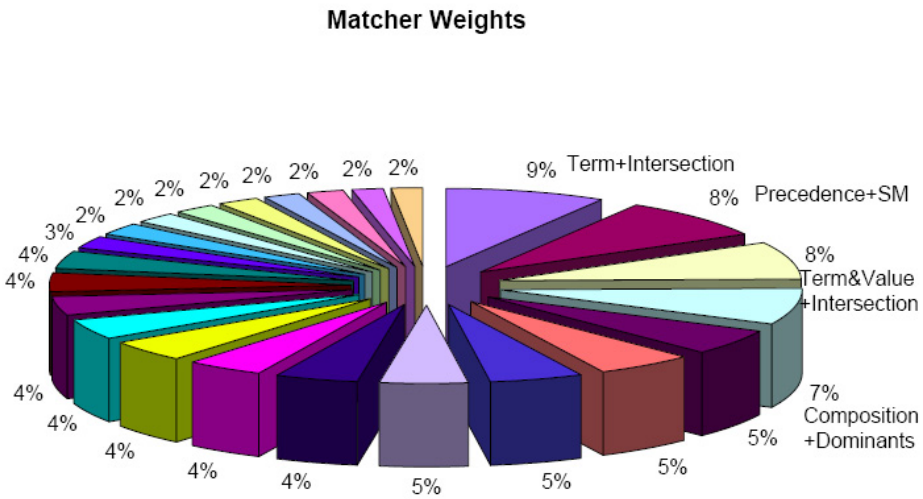
Example of SMB: cont.

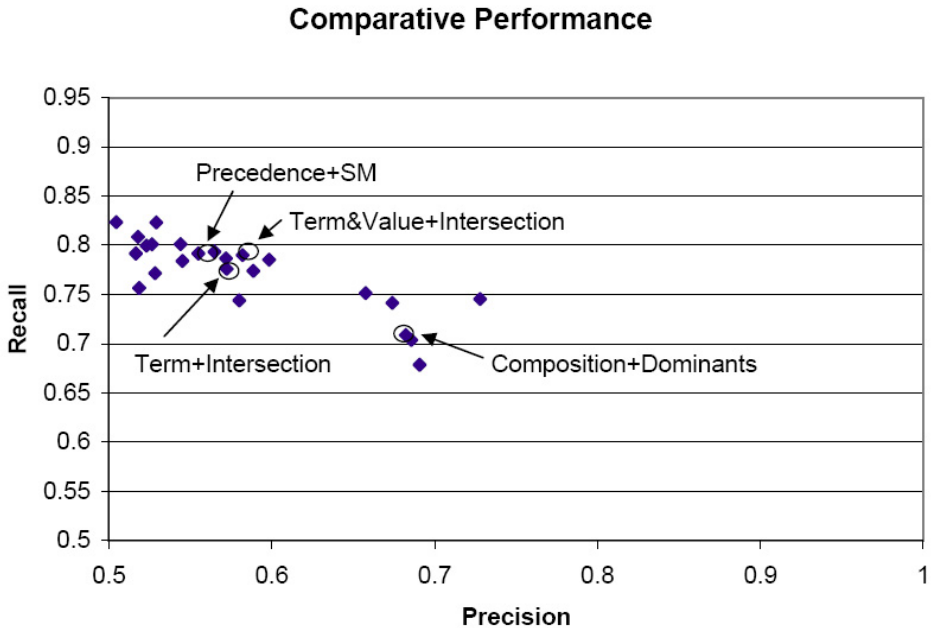
Example

- Given a new attribute pair (a, a') to be considered, each of the weak matchers contributes to the final decision such that its decision is weighted by its strength.
- If the final decision is positive, the given attribute pair is classified as an attribute correspondence. If not, it will be classified as incorrect.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Constructing Ensembles
AdaBoost
from AdaBoost to SMB
Top-K Schema Matchings

Decision Making of SMB





Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Constructing Ensembles

AdaBoost

from AdaBoost to SMB

Top-K Schema Matchings

Motivation

- Top-K schema matchings are intuitively defined as a ranked list of the best K schema matchings a matcher can generate.
- The formal definition is recursive, providing interesting insights into the behavior patterns of matchers.
- Top-K schema matchings play a pivotal role in managing uncertain schema matching.
- The effectively unlimited heterogeneity and ambiguity of data description suggests that in many cases an exact matching will not be identified as a best matching by any schema matcher.
- Top-K schema matchings are useful:
  - Creating a search space in uncertain settings.
  - Assigning probabilities in probabilistic schema matchings.
  - Improving the precision of matching results.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Definition

- Let $G = (X, Y, E)$ be an undirected bipartite graph with nodes representing attributes of two schemata and edges representing the degree of similarity between attributes.
- Assume a problem instance with a positive weight function $\varpi : E \rightarrow (0, 1]$ defined on edges.
- We are given a schema matcher and a similarity matrix M, $\varpi(i, j) = M_{i,j}$.
- G contains no edges with 0 weight.
- A matching σ is a subset of G's edges, $\sigma \subseteq E$.
- $\sigma \subseteq E$ is equivalent to $\sigma \in \Sigma$.
- The weight of a matching σ is $f(\sigma, M)$
- Given a constraint specification Γ, we consider only valid schema matchings in Σ_Γ.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Definition: cont.

- Top-K schema matchings is defined recursively.
- For $K = 1$, the K-th best matching σ_1^* is any maximum weight matching in G satisfying

$$\forall \sigma \subseteq E, f(\sigma, M) \leq f(\sigma_1^*, M).$$

- Let σ_i^* denote the i-th best matching, for any $i > 1$.
- Given the best $i-1$ matchings $\sigma_1^*, \sigma_2^*, \ldots, \sigma_{i-1}^*$, the i-th best matching σ_i^* is a matching of maximum weight over matchings that differ from each of $\sigma_1^*, \sigma_2^*, \ldots, \sigma_{i-1}^*$.
- Given top-K matchings, any matching $\sigma \subseteq E$ such that $\sigma \notin \{\sigma_1^*, \sigma_2^*, \ldots, \sigma_K^*\}$ satisfies

$$f(\sigma, M) \leq \min_{1 \leq j \leq k} f(\sigma_j^*, M) = f(\sigma_K^*, M).$$

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

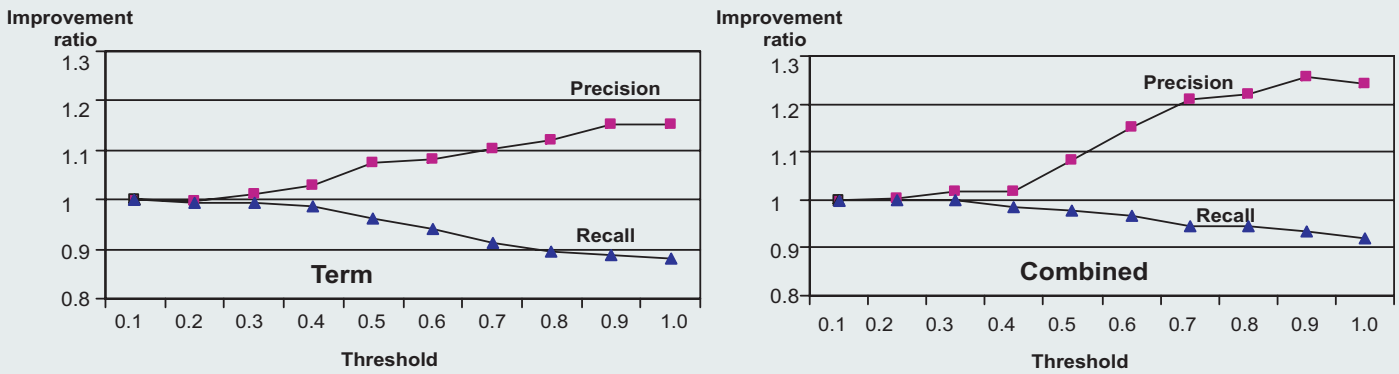
Intuitive interpretation

- Suppose an edge weight represents a matcher's belief in the correctness of an attribute correspondence, where a higher weight indicates greater confidence.
- When switching from the i-th best matching to the $(i+1)$ best matching, the matcher is forced to give up at least one attribute correspondence, while maintaining an overall high confidence in the matching.
- To do so, the matcher cedes an attribute correspondence in which it is less confident.
- Generating top-K matchings can be seen as a process by which a matcher iteratively abandons attribute correspondences in which it is less confident.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

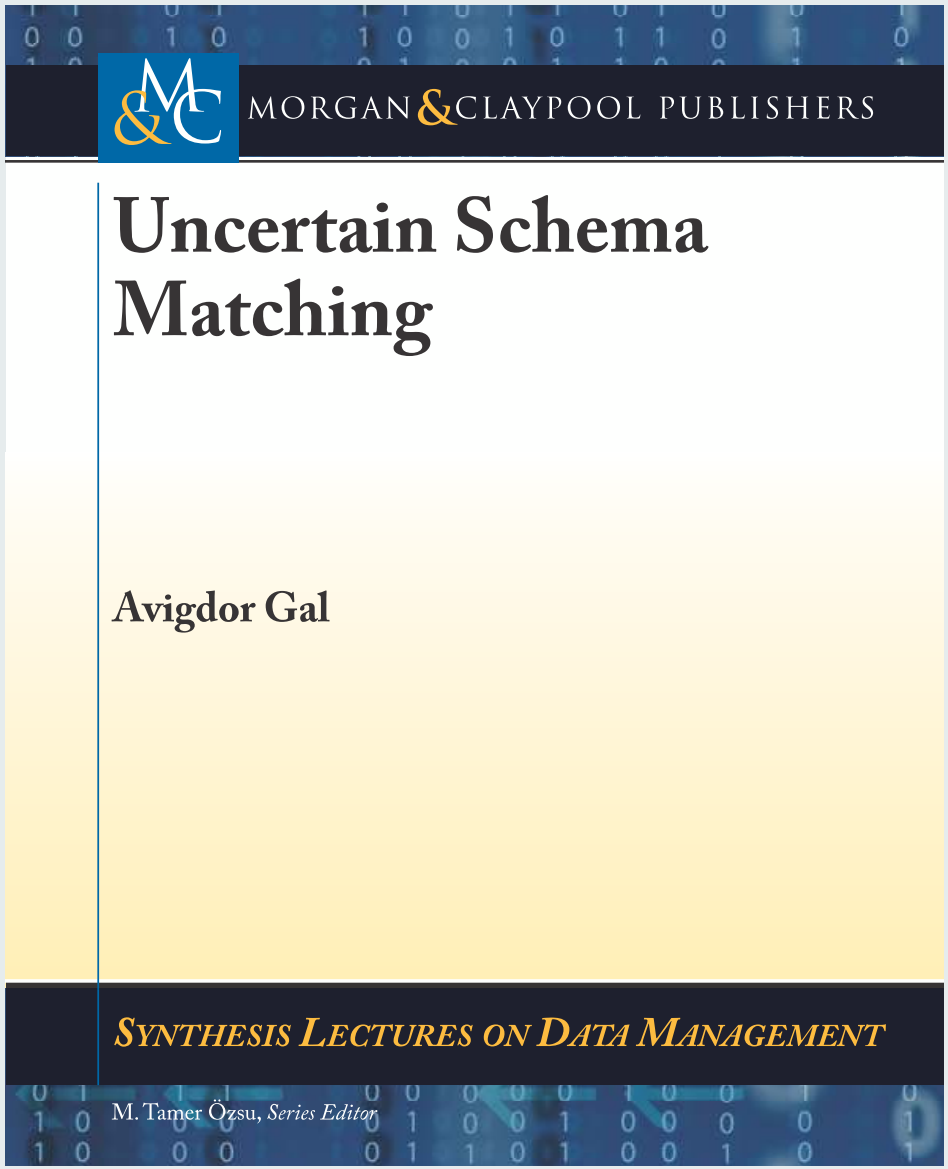
Impact of top-K Matching: an Empirical Example

Figure: Precision and recall for stability analysis with $K = 10$



Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Resources



Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

A. Algergawy.
Management of XML Data by Means of Schema Matching.

PhD thesis, Otto-von-Guericke University, 2010.

D. Barbosa, J. Freire, and A. Mendelzon.
Designing information-preserving mapping schemes for XML.
In *Proc. 31st Int. Conf. on Very Large Data Bases*, pages 109–120, 2005.

C. Batini, M. Lenzerini, and S. Navathe.
A comparative analysis of methodologies for database schema integration.
ACM Computing Surveys, 18(4):323–364, Dec. 1986.

M. Benerecetti, P. Bouquet, and S. Zanobini.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

Soundness of schema matching methods.
In *Proceedings of the 2nd European Semantic Web Conference*, pages 211–225, 2005.

J. Berlin and A. Motro.
Autoplex: Automated discovery of content for virtual databases.
In *Proc. Int. Conf. on Cooperative Information Systems*, pages 108–122. Springer, 2001.

P. Bernstein, S. Melnik, M. Petropoulos, and C. Quix.
Industrial-strength schema matching.
SIGMOD Record, 33(4):38–43, 2004.

P. Bohannon, W. Fan, M. Flaster, and P. Narayan.
Information preserving XML schema embedding.
In *Proc. 31st Int. Conf. on Very Large Data Bases*, pages 85–96, 2005.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

R. Cheng, J. Gong, and D. Cheung.
Managing uncertainty of XML schema matching.
In *Proc. 26th Int. Conf. on Data Engineering*, pages 297–308, 2010.

H. Do and E. Rahm.
COMA - a system for flexible combination of schema matching approaches.
In *Proc. 28th Int. Conf. on Very Large Data Bases*, pages 610–621, 2002.

A. Doan, P. Domingos, and A. Halevy.
Reconciling schemas of disparate data sources: A machine-learning approach.
In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 509–520, 2001.

C. Domshlak, A. Gal, and H. Roitman.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

Rank aggregation for automatic schema matching.
IEEE Trans. Knowl. and Data Eng., 19(4):538–553, 2007.

X. Dong, A. Halevy, and C. Yu.
Data integration with uncertainty.
In *Proc. 33rd Int. Conf. on Very Large Data Bases*, pages 687–698, 2007.

F. Duchateau, Z. Bellahsene, and R. Coletta.
A flexible approach for planning schema matching algorithms.
In *Proc. Int. Conf. on Cooperative Information Systems*, pages 249–264, 2008.

D. Embley, D. Jackman, and L. Xu.
Attribute match discovery in information integration: Exploiting multiple facets of metadata.
Journal of Brazilian Computing Society, 8(2):32–43, 2002.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

W. Frakes and R. Baeza-Yates, editors.
Information Retrieval: Data Structures & Algorithms.
Prentice Hall, Englewood Cliffs, NJ 07632, 1992.

Y. Freund and R. Schapire.
A decision-theoretic generalization of on-line learning and an application to boosting.
Journal of Computer and System Sciences, 55(1):119–139, Aug. 1997.

Y. Freund and R. Schapire.
A short introduction to boosting, 1999.

A. Gal.
Managing uncertainty in schema matching with top-k schema mappings.
Journal of Data Semantics, 6:90–114, 2006.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi.
A framework for modeling and evaluating automatic semantic reconciliation.
VLDB J., 14(1):50–67, 2005.

A. Gal, M. Martinez, G. Simari, and V. Subrahmanian.
Aggregate query answering under uncertain schema mappings.
In *Proc. 25th Int. Conf. on Data Engineering*, pages 940–951, 2009.

A. Gal, G. Modica, H. Jamil, and A. Eyal.
Automatic ontology matching using application semantics.
AI Magazine, 26(1):21–32, 2005.

A. Gal and T. Sagi.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Tuning the ensemble selection process of schema matchers.

Information Systems, 35(8):845–859, 2010.

Z. Galil.
Efficient algorithms for finding maximum matching in graphs.
ACM Comput. Surv., 18(1):23–38, Mar. 1986.

F. Giunchiglia, P. Shvaiko, and M. Yatskevich.
Semantic schema matching.
In *Proc. Int. Conf. on Cooperative Information Systems*, pages 347–365, 2005.

T. Green and V. Tannen.
Models for incomplete and probabilistic information.
Q. Bull. IEEE TC on Data Eng., 29(1):17–24, 2006.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

D. Gusfield and R. Irving.
The Stable Marriage Problem: Structure and Algorithms.
MIT Press, Cambridge, MA, 1989.

J. Halpern.
Reasoning About Uncertainty.
MIT Press, 2003.

B. He and K.-C. Chang.
Making holistic schema matching robust: an ensemble approach.
In *Proc. 11th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 429–438, 2005.

B. He and K. C.-C. Chang.
Statistical schema matching across Web query interfaces.
In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, pages 217–228, 2003.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

R. Hull.
Relative information capacity of simple relational database schemata.
SIAM J. Comput., 15(3):856–886, 1986.

Y. Lee, M. Sayyadian, A. Doan, and A. Rosenthal.
eTuner: tuning schema matching software using synthetic scenarios.
VLDB J., 16(1):97–122, 2007.

J. Madhavan, P. Bernstein, A. Doan, and A. Halevy.
Corpus-based schema matching.
In *Proc. 21st Int. Conf. on Data Engineering*, pages 57–68, 2005.

J. Madhavan, P. Bernstein, and E. Rahm.
Generic schema matching with Cupid.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

In *Proc. 27th Int. Conf. on Very Large Data Bases*, pages 49–58, 2001.

M. Magnani, N. Rizopoulos, P. McBrien, and D. Montesi.
Schema integration based on uncertain semantic mappings.

In *Proc. 24th Int. Conf. on Conceptual Modeling*, pages 31–46, 2005.

A. Marie and A. Gal.
Managing uncertainty in schema matcher ensembles.
In H. Prade and V. Subrahmanian, editors, *Scalable Uncertainty Management, First International Conference*, pages 60–73, 2007.

A. Marie and A. Gal.
On the stable marriage of maximum weight royal couples.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources

In *Proceedings of AAAI Workshop on Information Integration on the Web*, 2007.

A. Marie and A. Gal.
Boosting schema matchers.
In *Proc. Int. Conf. on Cooperative Information Systems*, pages 283–300, 2008.

S. Melnik, H. Garcia-Molina, and E. Rahm.
Similarity flooding: A versatile graph matching algorithm and its application to schema matching.
In *Proc. 18th Int. Conf. on Data Engineering*, pages 117–140, 2002.

E. Mena, V. Kashayap, A. Illarramendi, and A. Sheth.
Imprecise answers in distributed environments: Estimation of information loss for multi-ontological based query processing.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

Int. J. Cooperative Information Syst., 9(4):403–425, 2000.

R. Miller, Y. Ioannidis, and R. Ramakrishnan.
The use of information capacity in schema integration and translation.
In *Proc. 19th Int. Conf. on Very Large Data Bases*, pages 120–133, 1993.

G. Modica, A. Gal, and H. Jamil.
The use of machine-generated ontologies in dynamic information seeking.
In *Proc. Int. Conf. on Cooperative Information Systems*, pages 433–448, 2001.

P. Mork, A. Rosenthal, L. Seligman, J. Korb, and K. Samuel.
Integration workbench: Integrating schema integration tools.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

In *Proc. 22nd Int. Conf. on Data Engineering Workshops*, page 3, 2006.

H. Nottelmann and U. Straccia.
sPLMap: A probabilistic approach to schema matching.
In *Proc. Advances in Information Retrieval, 27th European Conference on IR Research, ECIR 2005*, pages 81–95, 2005.

S. Ross.
A First Course in Probability.
Prentice Hall, fifth edition, 1997.

R. Schapire.
The strength of weak learnability.
Machine Learning, 5:197–227, 1990.

W. Su, J. Wang, and F. Lochovsky.
A holistic schema matching for Web query interfaces.

Motivating Example
Introduction
Models of Uncertainty
Modeling Uncertain Schema Matching
Assessing Matching Quality
Schema Matcher Ensembles
Top-K Schema Matchings
Resources

In *Advances in Database Technology, Proc. 10th Int. Conf. on Extending Database Technology*, pages 77–94, 2006.

L. Zadeh.
Fuzzy sets.
Information and Control, 8:338–353, 1965.

Motivating Example

Introduction

Models of Uncertainty

Modeling Uncertain Schema Matching

Assessing Matching Quality

Schema Matcher Ensembles

Top-K Schema Matchings

Resources