A probabilistic approach to schema matching

Henrik Nottelmann[1] and Umberto Straccia[2]

[1] Institute of Informatics and Interactive Systems, University of Duisburg-Essen, 47048 Duisburg, Germany, nottelmann@uni-duisburg.de
[2] ISTI-CNR, Via G. Moruzzi 1, 56124 Pisa, Italy, straccia@isti.cnr.it
Number: TR-2004-xx

Abstract. This paper introduces the first formal framework for learning mappings between heterogeneous schemas, which is based on probabilistic logics. This task, also called "schema matching", is a crucial step in integrating heterogeneous collections. As schemas may have different granularities, and as schema attributes do not always match precisely, a general-purpose schema mapping approach requires support for uncertain mappings, and mappings have to be learned automatically. The framework combines different classifiers for finding suitable mapping candidates (together with their weights), and selects that set of mapping rules which is the most likely one. Finally, the framework with different variants has been evaluated on two different data sets.

ACM Categories and Subject Descriptors: H.3.5 **[Online Information Services]**: *Data sharing*;
Keywords: Schema matching, probabilistic logic, machine learning

1 Introduction

Federated digital libraries integrate a large number of legacy libraries and give users the impression of one coherent, homogeneous library. These libraries use different schemas (called source schemas). As users cannot deal efficiently with this semantic heterogeneity, they only see one system-wide or personalized target (or global) schema, which is defined ontologically and independent from the libraries. Then, queries are transformed from the target (global) schema into the source schemas, and documents vice versa (which is out of the scope of this paper).

Our framework uses probabilistic logics for describing schema mappings. In contrast to most of the approaches available so far, this allows dealing with schemas of different granularity. If the target schema contains the two attributes "author" and "editor", and the source schema only the more general attribute "creator", this source attribute cannot be mapped onto "author" precisely but only with a specific probability. Systems with purely deterministic mappings fail in such settings.

Here, we focus on learning these schemas using documents in both schemas, but not necessarily the same documents. As a by-product, we also compute a theoretically founded measurement for the quality of a mapping.

For schemas, we adopt the document model presented in [7] with only slight modifications. Like in database systems, data types with comparison operators are explicitly

modelled. However, vagueness of query formulations is one of the key concepts of Information Retrieval. Thus, it is crucial that comparison operators have a probabilistic interpretation. Vagueness is required e.g. when a user is uncertain about the exact publication year of a document or the spelling of an author name. These comparison operators are called often called "vague predicates", we will use the term "operator" later. For a specific attribute value the vague predicate yields an estimate of the probability that the condition is fulfilled from the user's point of view — instead of a Boolean value as in DB systems. The schema mapping rules also cover the problem of converting one query condition, a triple of attribute name, operator and comparison value, in another schema, where potentially also the operator or the comparison value has to be modified.

The paper is structured as follows: The next section introduces a formal framework for schema mapping, based on probabilistic logics. Section 3 presents a theoretically founded approach for learning these schema mappings, based on the combination of different classifiers. This approach is evaluated on two different test beds in section 4. Then, section 5 describes how this work is related to other approaches. The last section summarises this paper and gives an outlook over future work.

2 Formal framework for schema mapping

This section introduces a formal, logics-based framework for schema mapping. It shares a lot of ideas from other approaches, e.g. [5], but is different as it is the first one which also takes data types, predicates and query mapping into consideration. It is also the first framework, which is able to cope with the intrinsic uncertainty of the mapping process. The framework is based on probabilistic Datalog [8], for which tools are available.

2.1 Probabilistic Datalog

Probabilistic Datalog (pDatalog for short) is an extension to Datalog, a variant of predicate logic based on function-free Horn clauses. Negation is allowed, but its use is limited to achieve a correct and complete model. In pDatalog every fact or rule has a probabilistic weight $0 < \alpha \leq 1$ attached, prefixed to the fact or rule:

$$\alpha\, A \leftarrow L_1, \ldots, L_n\ .$$

Here, A denotes an atom (in the rule head), and $L_1, \ldots, L_n$ ($n \geq 0$) are literals, i.e atoms or negated atoms (the sub goals of the rule body). A weight $\alpha = 1$ can be omitted. Each fact and rule can only appear once in the program, to avoid inconsistencies. The intended meaning of a rule αr is that "the probability that any instantiation of rule r is true is α". The following example pDatalog program expresses the fact that 50% of all persons are male:

$$\begin{aligned} \texttt{person(mary)} &\leftarrow \\ 0.8\ \texttt{person(ed)} &\leftarrow \\ 0.5\ \texttt{male(X)} &\leftarrow \texttt{person(X)} \end{aligned}$$

Thus, $Pr(\texttt{male(mary)}) = 0.5$, and $Pr(\texttt{male(ed)}) = 0.8 \times 0.5 = 0.4$. Formally, an interpretation[3] in pDatalog is a tuple $I = (\Delta, \mathcal{W}, \mu)$. Here, $\mathcal{W}$ denotes a possible world

[3] In probabilistic Datalog, only Herbrand interpretations are considered.

(the instantiation of a the deterministic part of a pDatalog program plus a subset of the probabilistic part, where all probabilities are removed in the latter). In addition, μ is a probability distribution over $\mathcal{W}$. A model is an interpretation where for every world $w \in \mathcal{W}$ and every variable valuation v (a function mapping variables onto values from the domain, where $v(L)$ denotes the ground literal obtained by applying the v on L) the following conditions hold (where $L, L_1, \ldots, L_n$ are literals, r is a rule, and $0 < \alpha \leq 1$ is a weight):

$$(I,w,v) \models L \text{ iff } v(L) \in w,$$
$$(I,w,v) \models L \leftarrow L_1,\ldots,L_n \text{ iff whenever } (I,w,v) \models L_1,\ldots,L_n, \text{ then } (I,w,v) \models L.$$
$$(I,w,v) \models \alpha r \text{ iff } \mu(\{w' \in \mathcal{W} | (I,w',v) \models r\}) = \alpha\,.$$

In the remainder, we use an equivalent notion of interpretation. Here, an interpretation $I = (\Delta, \cdot^I)$ is a pair of the domain Δ and a function $\cdot^I$ which maps every ground atom onto a probability in $[0,1]$. The weights are determined based on an general independence assumption, but can consider ground atoms that are pair wise disjoint.

With abuse of notation, we typically do not distinguish between a relation R (an n-ary predicate) and the relation instance R^I (all ground instances of R w. r. t. I).

2.2 Data types

We first assume a finite set $\mathbf{D}$ of elementary data types. The domain $dom(d)$ for a data type $d \in \mathbf{D}$ defines the set of possible values for d. Examples are `Text` (for English text), `Name` (person names, e.g. "John Doe"), `Year` (four digit year numbers, e.g. "2004") or `DateISO8601` for the ISO 8601 format of dates (e.g. "2004-12-31"). We further use a set $\mathbf{O}$ of operators (sometimes also called "data type predicates"). An operator is a binary relation $o \subseteq dom(d_1(o)) \times dom(d_2(o))$ defined on two data types $d_1(o), d_2(o) \in \mathbf{D}$, e.g. `contains` for text (searching for stemmed terms), $>$ or $=$ for years, or `sounds-like` for names. The operator relations have a probabilistic interpretation (which is the probability that the first value matches the second one) for supporting vague queries. In our scenario, $\mathbf{D}$ contains the data type `DOCID` (the set of all document ids); only the identity operator $\mathtt{id}_{\mathtt{DOCID}}$ is defined on it.

As we want to use variables for operators, we use a bijective mapping between operators $o \in \mathbf{O}$ and new constants $\hat{o} \in \widehat{\mathbf{O}}$ for a set of constants $\widehat{\mathbf{O}}$. Then, these operators are combined in a ternary predicate `op`:

$$\mathtt{op} = \bigcup_{o \in \mathbf{O}} \{\hat{o}\} \times o\,. \tag{1}$$

Again, we do not explicitly distinguish between the operators o and their constants $\hat{o}$, and use the former notation for both of them. In addition, we use a predicate `conv` for value conversion between operators:

$$\begin{aligned}\mathtt{conv}^I \subseteq \textstyle\bigcup_{o_1,o_2 \in \mathbf{O}} & \{\hat{o_1}\} \times dom(d_1(o_1)) \times dom(d_2(o_1)) \times \\ & \{\hat{o_2}\} \times dom(d_1(o_2)) \times dom(d_2(o_2))\,.\end{aligned} \tag{2}$$

The informal meaning of $\mathtt{conv}(O,X,Y,O',X',Y')$ is that $\mathtt{op}(O,X,Y)$ can be transformed into $\mathtt{op}(O',X',Y')$. Also `conv` can be uncertain, where the weight denotes the

probability that this is a correct conversion. For example, `conv` may contain the tuples for the data types `Year2` (2-digit year numbers), `Year4` (4-digit year numbers), *FirstName* (only first names) and `Name` (complete names):

$$(\mathtt{id}_{\mathtt{Year2}}, \text{“}\mathtt{04}\text{”}, \text{“}\mathtt{04}\text{”}, \mathtt{id}_{\mathtt{Year4}}, \text{“}\mathtt{2004}\text{”}, \text{“}\mathtt{2004}\text{”}) \ ,$$
$$(\geq_{\mathtt{Year2}}, \text{“}\mathtt{04}\text{”}, \text{“}\mathtt{06}\text{”}, >_{\mathtt{Year4}}, \text{“}\mathtt{2005}\text{”}, \text{“}\mathtt{2005}\text{”}) \ ,$$
$$(\mathtt{id}_{\mathtt{FirstName}}, \text{“}\mathtt{John}\text{”}, \text{“}\mathtt{John}\text{”}, \mathtt{id}_{\mathtt{Name}}, \text{“}\mathtt{John\ Doe}\text{”}, \text{“}\mathtt{John\ Doe}\text{”}) \text{ with probability} < 1 \ .$$

2.3 Schemas and schema mappings

A schema $\mathbf{R} = \langle R_1, \ldots, R_n \rangle$ consists of a non-empty finite tuple of binary relation symbols. Each relation symbol R_i has a data type $d_{R_i} \in \mathbf{D}$. Then, for a (potentially uncertain) interpretation I, a schema instance is a tuple $\mathbf{R}^I = \langle R_1^I, \ldots, R_n^I \rangle$, where each relation symbol R_i is mapped onto a relation instance with the correct data types:

$$R_i \subseteq \mathtt{DOCID} \times dom(d_{R_i}) \ . \tag{3}$$

Informally, this is the relational model of linear schemas with multivalued schema attributes. Each attribute is modelled as a binary relation, which stores pairs of a document id and a value for that attribute.

Schema mappings follow the GaV approach [11]: A mapping is a tuple $\mathcal{M} = (\mathbf{T}, \mathbf{S}, \Sigma)$, where $\mathbf{T}$ denotes the target (global) schema and $\mathbf{S}$ the source (local) schema with no relation symbol in common, and Σ is a finite set of mapping constraints (pDatalog rules) of one of the forms (T_j and S_i are target and source attributes, respectively):

$$\alpha_{j,i} \ T_j(D,X) \leftarrow S_i(D,X_1), \mathtt{conv}(id_{d_{T_j}}, X, X, id_{d_{S_i}}, X_1, X_1) \tag{4}$$
$$\mathtt{op}(O,X,V) \leftarrow \mathtt{conv}(O,X,V,O_1,X_1,V_1), \mathtt{op}(O_1,X_1,V_1) \ . \tag{5}$$

For a schema mapping instance of a mapping $\mathcal{M} = (\mathbf{T}, \mathbf{S}, \Sigma)$ and a fixed interpretation I for $\mathbf{S}$, an interpretation J for $\mathbf{T}$ is a solution for I under $\mathcal{M}$ if and only if $\langle J, I \rangle$ (the combined interpretation over $\mathbf{T}$ and $\mathbf{S}$) satisfies Σ. The minimum solution is denoted by $J(I,\Sigma)$, the corresponding relation instance with $\mathbf{T}(I,\Sigma)$ (which is also called a minimum solution).

2.4 Queries and query transformation

A query q is a set of pDatalog rules with common head which define a unary predicate q with $q^I \subseteq \mathtt{DOCID}$. The literals of these rules refer to the relation symbols defined in $\mathbf{R} \cup \{\mathtt{op}\}$. The set $q^{\mathbf{R}^I}$ of answers for query q with respect to $\mathbf{R}^I$ contains exactly all the document ids which satisfy the query.

Given a schema mapping $\mathcal{M}$ and a source schema instance $\mathbf{S}^I$, the set $q^{\mathcal{M},\mathbf{S}^I}$ of certain answers to a query q (over $\mathbf{T} \cup \{\mathtt{op}\}$) with respect to $\mathcal{M}$ and $\mathbf{S}^I$ is exactly the set of answers for that query w. r. t. the minimum solution $\mathbf{T}(I,\Sigma)$:

$$q^{\mathcal{M},\mathbf{S}^I} = q^{\mathbf{T}(I,\Sigma)} \ . \tag{6}$$

In this paper, we are interested in correct (i.e. sound) reformulations. A query q' is a correct reformulation of a query q if we have $q'^{\mathbf{S}^I} \subseteq q^{\mathcal{M},\mathbf{S}^I}$ for every interpretation I. The tuples $t \in q'^{\mathbf{S}^I}$ then are certain answers, but q' does not necessary return all certain answers. This subset-property allows for handling cases in which no exact query transformation is possible. For example, consider the following scenario:

$$\begin{aligned}
\mathbf{T} &= \langle \texttt{year2} \rangle \,, \\
\mathbf{S} &= \langle \texttt{year4} \rangle \,, \\
\Sigma &= \{ \texttt{year2}(\texttt{D},\texttt{X}) \leftarrow \texttt{year4}(\texttt{D},\texttt{X}'), \texttt{conv}(\texttt{id}_{\texttt{Year2}},\texttt{X},\texttt{X},\texttt{id}_{\texttt{Year4}},\texttt{X}',\texttt{X}') \} \,, \\
\mathbf{S}^I &= \{(1,2000),(2,2001),(3,1990)\} \,, \\
\mathbf{T}^{J_1} &= \{(1,00),(2,01),(3,90),(4,02)\} \,, \\
\mathbf{T}^{J_2} &= \{(1,00),(2,01),(3,90),(5,04),(6,99)\} \,, \\
\mathbf{T}^{J_3} = \mathbf{T}(I,\Sigma) &= \{(1,00),(2,01),(3,90)\} \,.
\end{aligned}$$

Then, for source schema instance $\mathbf{S}^I$, both target schema instances $\mathbf{T}^{J_1}$ and $\mathbf{T}^{J_2}$ are solutions, and $\mathbf{T}^{J_3} = \mathbf{T}(I,\Sigma)$ is the minimal solution. Given the query $\texttt{q}(\texttt{X}) := \texttt{year2}(\texttt{D},\texttt{X}) \wedge \texttt{op}(\geq,\texttt{X},00)$ [4], the certain answers are $\texttt{q}^{\mathcal{M},\mathbf{S}^I} = \{(1,00),(2,01)\}$. Then, $\texttt{q1}(\texttt{X}) := \texttt{year4}(\texttt{D},\texttt{X}) \wedge \texttt{op}(\geq,\texttt{X},2000)$ and $\texttt{q2}(\texttt{X}) := \texttt{year4}(\texttt{D},\texttt{X}) \wedge \texttt{op}(>,\texttt{X},2000)$ are correct reformulations of $\texttt{q}$. The query $\texttt{q3}(\texttt{X}) := \texttt{year4}(\texttt{D},\texttt{X})$ is not even a correct reformulation, as:

$$\texttt{q3}^{\mathbf{S}^I} = \{(1,00),(2,01),(3,90)\} \not\subseteq \texttt{q}^{\mathcal{M},\mathbf{S}^I} \,.$$

3 Learning schema mappings

This paper only deals with learning schema mappings, i.e. finding associations between attributes. The assumption is that a set of data types **D** and a set of operators **O** with the corresponding relations op and conv are both already given. Learning schema mapping consists of three steps: The quality of potential schema mappings (set of rules) has to be estimated, the "best" schema mapping is selected, and, finally, the weights for rules in the selected schema mapping have to be estimated.

3.1 Estimating the quality of a schema mapping

For two schemas $\mathbf{T} = \langle T_1,\ldots,T_t \rangle$ and $\mathbf{S} = \langle S_1,\ldots,S_s \rangle$ and two interpretations I for **S** and J for **T**, the goal is to find a suitable set Σ of mapping constraints. In many cases, there is no correspondence between the tuples in both instances, so that no non-trivial mapping $\Sigma \supset \emptyset$ exists. Thus, the goal is to find the "best" set of mapping constraints Σ which maximizes the probability $Pr(\Sigma,J,I)$ that the tuples in the minimum solution $\mathbf{T}(I,\Sigma)$ under $\mathcal{M} = (\mathbf{T},\mathbf{S},\Sigma)$ and the tuples in **T** are plausible. Here, $\mathbf{T}(I,\Sigma)$ denotes a schema instance, and $T_j(I,\Sigma)$ the instance of relation T_j formed by the minimum solution. The set Σ can be partitioned into sets Σ_j with common head T_j, whose minimum solutions $T_j(I,\Sigma_j)$ only contain tuples for T_j:

$$\Sigma_t = \{r | r \in \Sigma, T_j \in \mathit{head}(r)\} \,, \tag{7}$$

$$\mathbf{T}(I,\Sigma) = \langle T_1(I,\Sigma_1),\ldots,T_t(I,\Sigma_t) \rangle \,. \tag{8}$$

[4] Select all documents published after 2000.

As a consequence, each target relation can be considered independently:

$$Pr(\Sigma, J, I) = \prod_{j=1}^{t} Pr(\Sigma_j, J, I) \ . \tag{9}$$

The instances $T_j(I, \Sigma_j)$ and T_j are plausible if the tuples in $T_j(I, \Sigma_j)$ are plausible values for T_j, and vice versa. Thus, $Pr(\Sigma_j, J, I)$ can be computed as:

$$Pr(\Sigma_j, J, I) = Pr(T_j | T_j(I, \Sigma_j)) \cdot Pr(T_j(I, \Sigma_j) | T_j) \tag{10}$$

$$= Pr(T_j(I, \Sigma_j) | T_j)^2 \cdot \frac{Pr(T_j)}{Pr(T_j(I, \Sigma_j))} \tag{11}$$

$$= Pr(T_j(I, \Sigma_j) | T_j)^2 \cdot \frac{|T_j|}{|T_j(I, \Sigma_j)|} \ . \tag{12}$$

As building blocks of Σ_j, we use the sets $\Sigma_{j,i}$ containing only one rule:

$$\Sigma_{j,i} = \{\alpha_{j,i}\ T_j(D, X) \leftarrow S_i(D, X), \mathtt{conv}(id_{d_{T_j}}, X, X, id_{d_{S_i}}, X', X')\} \ . \tag{13}$$

For s source relations and a fixed j, there are also s possible sets $\Sigma_{j,i}$, and $2^s - 1$ non-empty combinations (unions) of them, forming all possible non-trivial sets Σ_j. To simplify the notation, we set $S_i := T_j(I, \Sigma_{j,i})$ for the instance derived by applying the single rule (13). For computational simplification, we assume that S_{i_1} and S_{i_2} are disjoint for $i_1 \neq i_2$. Then, for $\Sigma_j = \bigcup_{k=1}^{r} \Sigma_{j,i_k}$ with indices $i_1, \ldots, i_r$, we obtain:

$$Pr(T_j(I, \Sigma_j) | T_j) = \sum_{k=1}^{r} Pr(S_{i_k} | T_j) \ . \tag{14}$$

Thus, the main task is to compute the $O(s \cdot t)$ probabilities $Pr(S_i | T_j)$.

3.2 Estimating the probability that a mapping rule is plausible

Computing the quality of a mapping requires the probability $Pr(S_i | T_j)$, while the rule weight is $\alpha_{j,i} = Pr(T_j | S_i)$. Both probabilities are estimated in a similar way. To ease handling of both directions, we use the letters A and B, respectively, and again identify A_i with A_i^I and B_j with B_j^J where necessary.

Similar to LSD [3], the probability $Pr(A_i | B_j)$ is estimated by combining different classifiers $C_1, \ldots C_n$. Each classifier C_k computes a weight $w(A_i, B_j, C_k)$, which has to be normalized and transformed into $Pr(A_i | B_j, C_k) = f(w(A_i, B_j, C_k))$, the classifier's approximation of $Pr(A_i | B_j)$. We employ different normalization functions f:

$$f_{id}(x) := x \ , \tag{15}$$

$$f_{sum}(x) := \frac{x}{\sum_{i'} w(A_{i'}, B_j, C_k)} \ , \tag{16}$$

$$f_{lin}(x) := c_0 + c_1 \cdot x \ , \tag{17}$$

$$f_{log}(x) := \frac{\exp(b_0 + b_1 \cdot x)}{1 + \exp(b_0 + b_1 \cdot x)} \ . \tag{18}$$

The functions f_{id}, f_{sum} and the logistic function f_{log} return values in $[0, 1]$. For the linear function, results below zero have to mapped onto zero, and results above one

have to be mapped onto one. The function f_{sum} ensures that each value is in $[0,1]$, and that the sum equals 1. Its biggest advantage is that is does not need parameters, which have to be learned. In contrast, the parameters of the linear and logistic function are learned by regression in a system-training phase. This phase is only required once, and their results can be used for learning arbitrary many schema mappings. Of course, normalization functions can be combined. In some cases it might be useful to bring the classifier weights in the same range (using f_{sum}), and the to apply another normalization function with parameters (e.g. the logistic function).

The final predictions $Pr(A_i|B_j,C)$ are then combined using the Total Probability Theorem, which results in a weighted sum:

$$Pr(A_i|B_j) \approx \sum_{k=1}^{n} Pr(A_i|B_j,C_k) \cdot Pr(C_k) \; . \tag{19}$$

The probability $Pr(C_k)$ describes the probability that we rely on the judgment of classifier C_k, which can for example be expressed by the confidence we have in that classifier. We simply use $Pr(C_k) = \frac{1}{n}$ for $1 \leq k \leq n$, i.e. the predictions are averaged.

Simple pDatalog rules can be used for computing the probabilities $Pr(A_i|B_j)$, as depicted in equation (19. With additional constants for the relations and classifiers, a binary predicate `alpha` for $Pr(A_i|B_j)$, a ternary predicate `alpha`$'$ for $Pr(A_i|B_j,C_k)$ and a unary predicate `alpha`$''$ for $Pr(C_k)$ with disjointness of the underlying events (i.e., $Pr(\texttt{alpha}''(c_{k_1}) \wedge \texttt{alpha}''(c_{k_2})) = 0$ for $k_1 \neq k_2$), the probabilities $Pr(A_i|B_j)$ can be computed using this pDatalog program:

$$\texttt{alpha}(a_i,b_j) \leftarrow \texttt{alpha}'(a_i,b_j,C),\texttt{alpha}''(C) \tag{20}$$

$$\alpha'_{i,j,k}\ \texttt{alpha}'(a_i,b_j,c_k) \leftarrow \forall 1 \leq k \leq n \tag{21}$$

$$\alpha''_{k}\ \texttt{alpha}''(c_k) \leftarrow \forall 1 \leq k \leq n \tag{22}$$

The weights $\alpha'_{i,j,k}$ are computed by the classifiers. Thus, in most cases, rule (21) can be replaced by additional rules.

3.3 Classifiers

Most classifiers require instances of both schemas. However, these instances do not need to describe the same objects. The instances should either be a complete collection, or a representative sample of it, e.g. acquired by query-based sampling [1]. Below, see a list of classifiers we considered.

Same attribute names. This binary classifier C_N returns a weight of 1 if and only if the two attributes have the same name, and 0 otherwise:

$$w(A_i,B_j,C_N) := \begin{cases} 1 & , \quad A_i = B_j, \\ 0 & , \quad \text{otherwise} \end{cases}$$

Introducing two new unary predicates `schema_a` (specifying if a constant corresponds to an attribute from schema **A**) and `schema_b` (similar for the schema **B**), the classifier can be expressed as:

$$\texttt{w}(\texttt{X},\texttt{X},\texttt{cn}) \leftarrow \texttt{schema\_a}(\texttt{A}),\texttt{schema\_b}(\texttt{X}) \; .$$

Exact tuples. This classifier C_E (for testing and evaluation) measures the fraction of the tuples in B_j which also occur in A_i:

$$w(A_i, B_j, C_E) := \frac{|A_i \cap B_j|}{|B_j|} .$$

Correct literals. This classifier C_L (suitable in particular for numbers, URLs and other facts) measures the fraction of the tuples in B_j where the data value (the second argument, without the document id) also occurs in any tuple in A_i:

$$w(A_i, B_j, C_L) := \frac{|\{s | s = (s_1, s_2) \in B_j, \exists t = (t_1, t_2) \in A_i . s_2 = t_2\}|}{|B_j|} .$$

kNN classifier. A popular classifier for text and facts is kNN [16]. For C_{kNN}, each attribute acts as a category, and training sets are formed for every tuple in A_l:

$$Train = \bigcup_{l=1}^{t} \{(A_l, t') | t' \in A_l\} . \tag{23}$$

A probabilistic variant of the scalar product is used for computing the similarity values. The values t and t' are considered as bags of words, and $Pr(w|A_i)$ and $Pr(w|B_j)$ are computed as the normalized frequencies of the words in the instances:

$$RSV(t, t') = \sum_{w \in t \cap t'} Pr(w|A_i) \cdot Pr(w|B_j) . \tag{24}$$

Naive Bayes text classifier. The classifier C_B uses a naive Bayes text classifier [16] for text content. Again, each attribute acts as a category, and attribute values are considered as bags of words (with normalised word frequencies as probability estimations). The final formula is:

$$w(A_i, B_j, C_B) = Pr(A_i) \cdot \sum_{x \in B_j} \prod_{w \in x} Pr(w|A_i) . \tag{25}$$

3.4 Estimating the weight of a rule

After a schema mapping (a set of rules) is learned, the weights $Pr(T_j|S_i)$ for these rules have to be computed. The probability $Pr(S_i|T_j)$ has already been computed for the quality estimation and, thus, can easily be transformed in the rule weight:

$$Pr(T_j|S_i) = Pr(S_i|T_j) \cdot \frac{Pr(T_j)}{Pr(S_i)} = Pr(S_i|T_j) \cdot \frac{|T_j|}{|S_i|} . \tag{26}$$

The drawback here is that the resulting rule weight might be greater than one in the case of $|T_j| > |S_i|$.

This completes the schema mapping learning process.

4 Experiments for learning schema mappings

This chapter describes the experiments conducted so far for evaluation the presented learning approach.

4.1 Evaluation setup

This section describes the test sets (source and target instances) and the classifiers used for the experiments. It also introduces different effectiveness measurements for evaluating the learned schema mappings (error, precision, recall). Experiments were performed on two different test beds [5]:

- BIBDB contains over 3,000 BibTeX entries about information retrieval and related areas. The documents are available both in BibTeX (source schema) and in the standard schema from the project MIND (target schema), derived from BibTeX via simple rules. Both schemas share a large amount of common attribute names.
- LOC is an Open Archive collection of the Library of Congress with about 1,700 documents, available in MARC 21 (source schema) and in Dublin Core (target schema). MARC 21 has a higher granularity as DC, thus a lot of DC attribute values are the concatenation of several MARC 21 attributes. Both schemas use a completely different name scheme, thus they do not have attribute names in common.

Each collection is split randomly into four sub-collections of approximately the same size. The first sub-collection is always used for learning the parameters of the normalization functions (same documents in both schemas). The second sub-collection is used as source instance for learning the rules, and the third sub-collection is used as the target instance. Finally, the fourth sub-collection is employed for evaluating the learned rules (for both instances, i.e. we evaluate on parallel corpora).

Each of classifiers introduced in section 3.3 are used alone, plus the combinations $C_{kNN}+C_B+C_L$ and $C_{kNN}+C_B+C_L+C_N$. The three normalization functions from section 3.2 (f_{sum}, f_{minmax} and f_{id}) are used; in every experiment, every classifier used the same normalization function.

The probability of a tuple t in the given target instance $\mathbf{T}_j^J$ is denoted by $Pr(t|T_j)$. Often the target instance only contains deterministic data, then we have $Pr(t|T_j) \in \{0,1\}$. Similarly, $Pr(t|T_j(I,\Sigma_j)) \in [0,1]$ denotes the probability of tuple t w. r. t. the minimal solution of the given source instance and the learned schema mapping, i.e. by applying the schema mapping on the source instance. Rule application includes mapping the resulting tuple weights onto 0 or 1, respectively, in the case where a rule weight α outside $[0,1]$ (due to a wrong estimation) leads to a tuple weight which is less than zero or higher than one.

The error of the mapping is defined by:

$$E(\mathcal{M}) = \frac{1}{\sum_j |U_j|} \sum_j \sum_{t \in U_j} \left(Pr(t|T_j) - Pr(t|T_j(I,\Sigma_j))\right)^2 , \tag{27}$$

$$U_j = T_j \cup T_j(I,\Sigma_j) . \tag{28}$$

[5] http://faure.isti.cnr.it/~straccia/download/TestBeds/ecir05-exp.tar.gz

Here, the set U_j contains the union of the given target instance tuples and the tuples created by applying the mapping rules. For each of these tuples, the squared difference of the given weight $Pr(t|T_j)$ in the target instance and the computed weight $Pr(t|T_j(I,\Sigma_j))$ is computed. Furthermore, we evaluated if the learning approach computes the correct rules (neglecting the corresponding rule weights). Similar to the area of Information Retrieval, precision defines how many learned rules are correct, and how many correct rules are learned. So, let R_L denote the set of rules (without weights) returned by the learning algorithm, and R_A the set of rules (again without weights) which are the actual ones. Then

$$precision := \frac{|R_L \cap R_A|}{|R_L|}, \quad recall := \frac{|R_L \cap R_A|}{|R_A|} . \tag{29}$$

4.2 Results

In the experiments presented in this section, the learning steps are as follows:

1. Find the best schema mapping
   (a) Estimate the plausibility probabilities $Pr(S_i|T_j)$ for every $S_I \in \mathbf{S}$, $T_j \in \mathbf{T}$ using the classifiers.
   (b) For every target relation T_j and for every non-empty subset of schema mapping rules having T_j as head, estimate the probability $Pr(\Sigma_j, J, I)$.
   (c) Select the rule set Σ_j which maximizes the probability $Pr(\Sigma_j, J, I)$.
2. Estimate the weights $Pr(T_j|S_i)$ for the learned rules by converting $Pr(S_i|T_j)$, using equation (26).
3. Compute the error, precision and recall as described above.

The results depicted in the tables 1 and 2 show that the LOC collection is much harder as the schemas have different granularities, and both schemas do not have any attribute name in common. The error for the BIBDB collection can be quite low (below 0.1 for C_L), while the error is always above 0.5 for LOC. Precision is high for both collections, but higher for BIBDB. As the learner C_N cannot learn any rule for LOC (as both schemas use completely different attribute names), the precision is not defined. For the BIBDB collection, recall can be quite high (over 0.7 for the combined classifiers). For LOC, however, the best recall achieved is 0.4146. Averaged on both collections and all normalization functions, the error is minimized by the combination $C_{kNN} + C_B + C_L + C_N$ with an error of 0.4334, followed by $C_{kNN} + C_B + C_L$ and C_{kNN}. Not surprisingly, C_N and C_E performed worst (more than 42% worse than $C_{kNN} + C_B + C_L + C_N$). These results are replicated considering recall. Interestingly, C_E yields the highest precision with 0.4339, followed by C_L and $C_{kNN} + C_B + C_L + C_N$ (about 20% worse). The worst precision (0.5 on average is obtained by C_N (474% worse than C_E). This last result is due to the fact that this classifier does not work on the LOC collection (with no attribute names in common), but perfectly works on the BIBDB collection. Overall, combining classifiers can reduce the error and increase recall and precision. Averaged on both collections and all classifiers, the best normalization functions w. r. t. the error are $f_{lin} \circ f_{sum}$ (0.4740) and $f_{log} \circ f_{sum}$ (about 1% worse). Precision is maximized for f_{id} (0.7074), while recall is maximized for f_{sum} and $f_{lin} \circ f_{sum}$ (both

	f_{id}	f_{sum}	$f_{lin} \circ f_{sum}$	$f_{log} \circ f_{sum}$
C_E	0.8613 / 0.0%	0.3675 / -57.3%	0.3719 / -56.8%	0.3713 / -56.9%
C_L	0.4035 / 0.0%	0.0999 / -75.2%	0.1077 / -73.3%	0.0767 / -81.0%
C_N	0.2641 / 0.0%	0.2641 / 0.0%	0.2874 / 8.8%	0.2874 / 8.8%
C_{kNN}	0.3549 / 0.0%	0.2752 / -22.4%	0.2875 / -19.0%	0.2800 / -21.1%
C_B	0.9224 / 0.0%	0.2871 / -68.9%	0.3036 / -67.1%	0.3123 / -66.1%
C_{kNN}+C_B+C_L	0.4211 / 0.0%	0.1495 / -64.5%	0.1622 / -61.5%	0.1678 / -60.2%
C_{kNN}+C_B+C_L+C_N	0.3585 / 0.0%	0.1494 / -58.3%	0.1765 / -50.8%	0.1830 / -48.9%

(a) Error

	f_{id}	f_{sum}	$f_{lin} \circ f_{sum}$	$f_{log} \circ f_{sum}$
C_E	1.0000 / 0.0%	1.0000 / 0.0%	1.0000 / 0.0%	1.0000 / 0.0%
C_L	0.7778 / 0.0%	0.7778 / 0.0%	0.7778 / 0.0%	0.7000 / 10.0%
C_N	1.0000 / 0.0%	1.0000 / 0.0%	1.0000 / 0.0%	1.0000 / 0.0%
C_{kNN}	0.5000 / 0.0%	0.5000 / 0.0%	0.5000 / 0.0%	0.6250 / -25.0%
C_B	0.5000 / 0.0%	0.4167 / 16.7%	0.4167 / 16.7%	0.4444 / 11.1%
C_{kNN}+C_B+C_L	0.5714 / 0.0%	0.5000 / 12.5%	0.5000 / 12.5%	0.5714 / 0.0%
C_{kNN}+C_B+C_L+C_N	0.8182 / 0.0%	0.8182 / 0.0%	0.7500 / 8.3%	0.8182 / 0.0%

(b) Precision

	f_{id}	f_{sum}	$f_{lin} \circ f_{sum}$	$f_{log} \circ f_{sum}$
C_E	0.3636 / 0.0%	0.3636 / 0.0%	0.3636 / 0.0%	0.3636 / 0.0%
C_L	0.6364 / 0.0%	0.6364 / 0.0%	0.6364 / 0.0%	0.6364 / 0.0%
C_N	0.6364 / 0.0%	0.6364 / 0.0%	0.6364 / 0.0%	0.6364 / 0.0%
C_{kNN}	0.5455 / 0.0%	0.5455 / 0.0%	0.5455 / 0.0%	0.4545 / 16.7%
C_B	0.0909 / 0.0%	0.4545 / -400.0%	0.4545 / -400.0%	0.3636 / -300.0%
C_{kNN}+C_B+C_L	0.7273 / 0.0%	0.7273 / 0.0%	0.7273 / 0.0%	0.7273 / 0.0%
C_{kNN}+C_B+C_L+C_N	0.8182 / 0.0%	0.8182 / 0.0%	0.8182 / 0.0%	0.8182 / 0.0%

(c) Recall

Table 1. ST-Rule(ST) – BIBDB

0.4067). The experiments show that using the trivial normalization function f_{id} dramatically increases the error (35%) and recall (6%), but performs best w. r. t. the precision.

The best classifier/normalization function combination is $C_{kNN} + C_B + C_L$ with $f_{lin} \circ f_{sum}$ with an error of 0.3907. Best precision is yield for using C_E with any normalization function (virtually no difference on average). Recall is maximized for $C_{kNN} + C_B + C_L + C_N$ with f_{id} (surprisingly), followed by the other normalization functions for the classifier combination. Thus, it is useful to combine classifiers.

	f_{id}	f_{sum}	$f_{lin} \circ f_{sum}$	$f_{log} \circ f_{sum}$
C_E	0.7474 / 0.0%	0.7805 / 4.4%	0.7264 / -2.8%	0.6975 / -6.7%
C_L	0.7106 / 0.0%	0.7518 / 5.8%	0.7016 / -1.3%	0.7135 / 0.4%
C_N	1.0000 / 0.0%	1.0000 / 0.0%	1.0000 / 0.0%	1.0000 / 0.0%
C_{kNN}	0.5766 / 0.0%	0.6215 / 7.8%	0.5708 / -1.0%	0.5805 / 0.7%
C_B	0.9561 / 0.0%	0.8040 / -15.9%	0.6203 / -35.1%	0.6308 / -34.0%
C_{kNN}+C_B+C_L	0.6368 / 0.0%	0.6928 / 8.8%	0.6192 / -2.8%	0.6539 / 2.7%
C_{kNN}+C_B+C_L+C_N	0.6449 / 0.0%	0.6577 / 2.0%	0.6286 / -2.5%	0.6684 / 3.6%

(a) Error

	f_{id}	f_{sum}	$f_{lin} \circ f_{sum}$	$f_{log} \circ f_{sum}$
C_E	0.9000 / 0.0%	0.9000 / 0.0%	0.9000 / 0.0%	0.8462 / 6.0%
C_L	0.7273 / 0.0%	0.7273 / 0.0%	0.7273 / 0.0%	0.6667 / 8.3%
C_N	not defined	not defined	not defined	not defined
C_{kNN}	0.7083 / 0.0%	0.7083 / 0.0%	0.7083 / 0.0%	0.5833 / 17.6%
C_B	1.0000 / 0.0%	0.5000 / 50.0%	0.5000 / 50.0%	0.6667 / 33.3%
C_{kNN}+C_B+C_L	0.7000 / 0.0%	0.6471 / 7.6%	0.6471 / 7.6%	0.6667 / 4.8%
C_{kNN}+C_B+C_L+C_N	0.7000 / 0.0%	0.6471 / 7.6%	0.6471 / 7.6%	0.6667 / 4.8%

(b) Precision

	f_{id}	f_{sum}	$f_{lin} \circ f_{sum}$	$f_{log} \circ f_{sum}$
C_E	0.2195 / 0.0%	0.2195 / 0.0%	0.2195 / 0.0%	0.2683 / -22.2%
C_L	0.1951 / 0.0%	0.1951 / 0.0%	0.1951 / 0.0%	0.1951 / 0.0%
C_N	0.0000 / 0.0%	0.0000 / 0.0%	0.0000 / 0.0%	0.0000 / 0.0%
C_{kNN}	0.4146 / 0.0%	0.4146 / 0.0%	0.4146 / 0.0%	0.3415 / 17.6%
C_B	0.0244 / 0.0%	0.1463 / -500.0%	0.1463 / -500.0%	0.0976 / -300.0%
C_{kNN}+C_B+C_L	0.3415 / 0.0%	0.2683 / 21.4%	0.2683 / 21.4%	0.2439 / 28.6%
C_{kNN}+C_B+C_L+C_N	0.3415 / 0.0%	0.2683 / 21.4%	0.2683 / 21.4%	0.2439 / 28.6%

(c) Recall

Table 2. ST-Rule(ST) – LOC

As an illustrative example, in one of BIBDB runs, these two rules are returns for the target attribute `booktitle`:

$$0.51\ \texttt{standard\_booktitle}(\texttt{D},\texttt{X}) \leftarrow \texttt{BIBDB\_booktitle}(\texttt{D},\texttt{X}'),\ \texttt{conv}(\texttt{id}_{\texttt{Text}},\texttt{X},\texttt{X},\texttt{id}_{\texttt{Text}},\texttt{X}',\texttt{X}')$$

$$0.98\ \texttt{standard\_booktitle}(\texttt{D},\texttt{X}) \leftarrow \texttt{BIBDB\_journal}(\texttt{D},\texttt{X}'),\ \texttt{conv}(\texttt{id}_{\texttt{Text}},\texttt{X},\texttt{X},\texttt{id}_{\texttt{Text}},\texttt{X}',\texttt{X}')$$

Notice that, for instance, a query for `booktitle` is then converted into the source schema, using the above rules, by unfolding the query into two source queries (one for `booktitle`, the other for `journal`).

5 Related work

In the field of federated databases, two approaches are distinguished (see [11, 15]). In "local as view" (LaV), the source schemas are defined as views (mappings) over a fixed global schema. This makes it easy to add a new source, but query transformation has exponential time complexity. In contrast, the global schema is defined as a view over local schemas in the "global as view" (GaV) approach. Here, query transformation can be reduced to rule unfolding, but the adding of new sources might require to modify the global view. The GLaV approach [6] combines the advantages of both worlds. The global schema is specified ontologically and independent from the sources, the source schema models the documents returned by the source, and mappings are defined by logical rules between query expressions. We adopt the main GLaV idea of independent schemas, but use probabilistic GaV rules, and restrict the schema structure to binary relations (for attributes).

A general approach for learning rules (not only for schema mapping) is described in [12]. ILP (Inductive Logic Programming) is employed for learning rules, while PAC learning algorithm is used for learning the rule weights. The approach requires the same documents in both schemas ("parallel corpora"), which is infeasible in most environments. A second drawback is that it is based on exact match only.

Similar to our approach, the heuristic system LSD [3] for finding 1:1 matchings in XML documents uses a linear combination of the predictions of multiple base learners (classifiers). The combination weights are learned via regression on manually specified mappings between a small number of learning schemas. LSD has several extensions, e.g. iMAP [2] for complex matchings in relational databases and GLUE [4] for matching ontologies on the semantic web (which relies on joint probability distributions).

Information theory measures and graph matching is used in [10]. Graphs are constructed from the schemas, where the attributes form the nodes, labelled with the entropy of the attribute. All nodes are connected, the edges are labelled with the mutual information (correlation between two distributions). Both measures do not require any interpretation of the data, i.e. data type do not have to be considered. A distance measure is defined, and optimum graph matchings is applied for finding schema mappings.

A completely different approach is taken in MGS [9]. It aims at finding a "hidden model", a schema that probabilistically generates the observed schemas. A hidden model is a partition of the attribute space with a probability function of the partitions and their attributes. The first step finds cliques in the graph where two nodes (attributes) are connected if they are not occurring in the same schema. These cliques do not contradict the schemas. The problem of selecting those cliques which form a partitions is then reduced to a set-cover problem, and the probability functions are computed by maximum-likelihood. In a final step, χ^2 statistical testing is employed for finding sufficiently consistent models.

6 Conclusion and outlook

In this paper we introduced a formal GLaV-like framework for schema mappings, where the mappings are defined as uncertain rules in probabilistic Datalog. These schema

mapping rules do not only cover transforming data from one attribute into another, but can also be used for transforming query conditions (potentially also modifying the operator or the comparison value). Although the framework is based on logics, real-world documents and queries with a linear schema can easily be converted into the logical formalism.

We also presented an approach for learning schema mappings. Different classifiers are used for predicting the probability that tuples in a target relation are plausible for a source relation. Similar to LSD, these predictions are combined to an overall approximation of probability. From these probabilities, a probability that a set of such schema mapping rules is plausible is derived. Finally, the rule weights have to be computed. The evaluation shows that the system can be used in practice.

The results in this paper can be used in different ways:

1. Specific schema mapping services can be automatically built. Each schema mapping service has associated two schemas, and it is responsible for mapping between these two schemas. The mapping "function" should be learned automatically instead of being defined manually.
2. Peer-to-peer networks are dynamic scenarios where services can dynamically join and leave, so the system can–for each query–only consider the services, which are currently available. Using a decision-theoretic model as for the narrower task of resource selection, we have to find a quality measurement for a schema mapping service.

We mainly target at the information exchange problem: Two schemas are given, and an object instance in one schema is transformed into an instance of the other schema. Our mechanism could also be used for the problem of information integration: Given two source schemas, a mediated schema of them has to be created. A solution would be to build the union of both schemas, learn mapping rules, and remove useless attributes.

In future, more variants should be developed and evaluated to improve the quality of the learning mechanism. Additional classifiers could consider the data types of two attributes, could use a thesaurus for finding synonym attribute names, or could use other measures like KL-distance or mutual information. Instead of averaging the classifier predictions, the weights could be learned via regression. Odds or statistical significance tests could be employed for determining the best schema mapping.

In this work, the `conv` predicate is given. In environments with large numbers of data types, or a dynamically changing set of data types, learning the conversion predicate would be desirable, e.g. the conversion from centimeter to inch.

A more basic extension is the application onto ontologies. Instead of linear schemas, classification hierarchies are given. The task then is to map instances from one class onto classes in the other hierarchy.

7 Acknowledgements

This work is supported in part by ISTI-CNR (project "Distributed Search in the Semantic Web") and in part by the DFG (grant BIB47 DOuv 02-01, project "Pepper").

References

[1] J. Callan and M. Connell. Query-based sampling of text databases. *ACM Transactions on Information Systems*, 19(2):97–130, 2001.

[2] R. Dhamankar, Y. Lee, A. Doan, A. Halevy, and P. Domingos. iMAP: Discovering complex semantic matches between database schemas. In *SIGMOD 2004*, 2004.

[3] A. Doan, P. Domingos, and A. Y. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. In *SIGMOD Conference*, 2001.

[4] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the semantic web. 2004.

[5] R. Fagin, P. G. Kolaitis, W.-C. Tan, and L. Popa. Composing schema mappings: Second-order dependencies to the rescue. In *Proceedings PODS*, 2004.

[6] M. Friedman, A. Y. Levy, and T. D. Millstein. Navigational plans for data integration. In *Proceedings of 16th Natl Conf on Artificial Intelligence*, pages 67–73, 1999.

[7] N. Fuhr. Towards data abstraction in networked information retrieval systems. *Information Processing and Management*, 35(2):101–119, 1999.

[8] N. Fuhr. Probabilistic Datalog: Implementing logical information retrieval for advanced applications. *Journal of the American Society for Information Science*, 51(2):95–110, 2000.

[9] B. He and K. C.-C. Chang. Statistical schema matching across web query interfaces. In Papakonstantinou et al. [13].

[10] J. Kang and J. F. Naughton. On schema matching with opaque column names and data values. In Papakonstantinou et al. [13].

[11] M. Lenzerini. Data integration: a theoretical perspective. In *Proceedings of the 21st ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (PODS-02)*, pages 233–246. ACM Press, 2002.

[12] H. Nottelmann and N. Fuhr. Learning probabilistic Datalog rules for information classification and transformation. In Paques et al. [14], pages 387–394.

[13] Y. Papakonstantinou, A. Halevy, and Z. Ives, editors. *Proceedings SIGMOD 2003*, 2003.

[14] H. Paques, L. Liu, and D. Grossman, editors. *Proceedings of the 10th International Conference on Information and Knowledge Management*, New York, 2001. ACM.

[15] E. Rahm and P. A. Bernstein. A survey of approaches to automatic schema matching. *The VLDB Journal*, 10(4):334–350, 2001.

[16] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.