Entity-Attribute Resolution

Decision-theoretic symbol matching 102

Relevant Work

A Distance Based Approach to Entity Reconciliation in Hetergeneous Databases

- D. Dey and S. Sarkar (200109)
- Uses decision theory and user-solicited distances to determine yes/no entities in disparate databases are same.
- Assumes that semantic-heterogeneity of attributes has already been resolved.

Relevant Work

The Field Matching Problem: Algorithms and Applications

- A. Monge and C. Elkan (199608)
- Uses text matching algorithms to determine contents of attributes are same
- Does not work with symbolic data
- No subjective measure of success

Relevant Work

Automatic Ontology Mapping for Agent Communication

- F. Wiesman, N. Roos, P. Vogt (200207)
- Establishes mappings using a *joint attention set* based on word co-occurrence
- Cannot resolve symbolic data
- No metric for deciding whether translation was feasible

Relevant Work

Algorithms for Ontological Mediation

- E. Campbell and S. Shapiro (199803)
- “Common words have common meanings”
- Uses WordNet lexical ontology to translate between two ontologies.
- Does not operate on symbolic data

Formulation – Records

Defining a record (of an event)…

- Let $\mathbf{G}$ be a finite, discrete set of "global keys"
- Let $\mathbf{K}$ be a finite, discrete set of "prediction keys"
- Let $\mathbf{S}$ be a finite, discrete set of "symbols" that encode an "attribute-of-interest"
- A record $\mathbf{r}$ is a triple of the form $<\mathbf{g_i}, \mathbf{k_i}, \mathbf{s_i}>$ where $\mathbf{g_i} \in \mathbf{G}$, $\mathbf{k_i} \in \mathbf{K}$ and $\mathbf{s_i} \in \mathbf{S}$

Formulation - Metadatabase

Define a metadatabase as:

- Let **P(K)** denote the distribution of **K** over the domain of discourse
- Let **P(S)** denote the distribution of **S** over the domain of discourse.
- Let **C** be a set of conditional probabilities such that $\exists\, P(s_i|k_j)$ such that $P(s_i|k_j) \neq P(s_i)$
- A metadatabase **META** is the tuple…

<G,K,P(K),S,P(S),C>

Formulation – Global Keys

Elements of G are globally unique identifiers.

- For any two records $\mathbf{r}_i$ and $\mathbf{r}_j$ in a set of records, $\mathbf{R}$: $\mathbf{g}_i \neq \mathbf{g}_j$

Formulation - Database

A database is then defined as follows…

- A set of records $\mathbf{R} = \{r_1, \ldots, r_n\}$
- A database $\mathbf{D}$ is the tuple $\langle \mathbf{M}, \mathbf{R} \rangle$ where $\mathbf{M}$ is a metadatabase describing the the domain of discourse and $\mathbf{R}$ is a set of recorded observations of events in the domain of discourse.

Formulation - Agent

An agent is an entity that...

- Has knowledge about its environment
- is **active**
- is **autonomous**
- seeks **new information resources**
- **acts on this information** to increase utility

Formulation – Agent – Observation

We allow for the possibility that an agent's observations are not perfect

- This is represented using conditional probabilities of the form $\mathtt{P(s_o|s_a)}$
- Let **`DISCRIMINATOR`** be an $\mathtt{S} \times \mathtt{S}$ matrix encoding the conditional probabilities.

Formulation – Agent – Actions

An agent takes action based upon observation of objects or events in the domain-of-discourse.

- Let the actions (strategies) available to the agent be the set **ACTIONS** $= \{a_1, ..., a_n\}$

Formulation – Agent – Outcomes

Upon the execution of some action, an agent may experience any one of a set of finite outcomes.

- Let **`OUTCOMES`** be the finite, discrete set of possible results from the execution of some action.

Formulation – Agent – Payoffs

A payoff is what an agent experiences when executing a particular action on the stimulus of some event or object.

- Let `PAYOFFS` be a matrix of `S` $\times$ `OUTCOMES`
- A payoff `PAYOFF(a|s)` is some distribution over `OUTCOMES`.

Formulation – Agent – Information

An agent may have one or more information sources.

- Let **INFO** be a set of tuples of the form $\langle t_i, R_i \rangle$ where t_i is a translator that maps the attribute of interest of records in R_i into the symbol set S of the agent's metadatabase.

Formulation - Agent

An agent is thus formulated as the tuple of...

```
A = <META,
     DISCRIMINATOR,
     ACTIONS,
     OUTCOMES,
     PAYOFFS,
     INFO>
```

Formulation – Agent - Operations

We will need these functions for later computation:

- **`choice(A,s)`** – Agent A's optimal strategy for acting on observation of object of class s with certainty.
- **`action(A,`**s_o**`,`**s_a**`)`** – The distribution of outcomes of agent **`A`** executing **`choice(A,`**s_o**`)`** on an event/object of class s_a.

Formulation - Translation

- A translation function $\mathbf{t}$ is a mapping defined over $\mathbf{S_r} \times \mathbf{S_l}$: $\mathbf{t(S_r)} \rightarrow \mathbf{S_l}$
- Let $|\mathbf{S_r}| = \mathbf{m}$
- Let $|\mathbf{S_l}| = \mathbf{n}$
- The set of all translation functions is denoted $\mathbf{T_{S_r S_l}}$ and contains $\mathbf{m^n}$ translations.
- Translation is also defined on records where the attribute of interest is translated through $\mathbf{t}$.

Formulation - Translation

Some useful functions for later...

- **`default`**(S_r) – Constructs a default translator by mapping all elements of S_r to the unknown symbol.
- **`update`**(t, s_r, s_l) – Produces a new translator by associating the remote symbol s_r with the local symbol s_l

Statement

- Given an agent under local control $\mathtt{A_l}$
- Given a remote set of records $\mathtt{R_r}$ with $\mathtt{K_l \cap K_r \neq \{\}}$
- Find translation function $\mathtt{t}$ in $\mathtt{T_{s_l s_r}}$ such that $\Sigma\, \mathtt{E[action(A_l, t(r \in R_r))]}$ is maximal.

Solution – The Unknown

We begin the solution by augmenting $\mathbf{S_r}$ with a new symbol (**?**) that indicates a complete lack of knowledge.

Likewise, the **ACTIONS** set of $\mathbf{A_1}$ must also be augmented with a null strategy which is executed only on observation of the unknown.

Solution – E[Value] of a Record

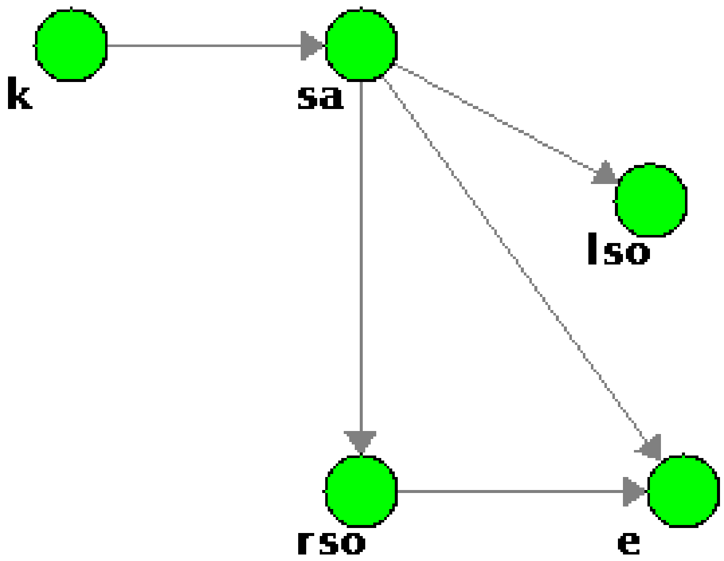
To compute the aggregate value of a translation, we first develop a function to determine the expected value of acting on an observation.

Solution – E[Value] of a Record

The process will consist of:

1. Apply translation: $\mathbf{r_t \leftarrow t(r)}$
2. Select the best strategy: **`choice(A,class(`r_t`))`**
3. Use a BBN to compute the distribution of **`class(`r_t`)`** over $\mathbf{S}$.
4. Compute the expected value of executing this strategy.

Solution – E[Value] of a Record



BBN for computing distribution of class $\mathbf{s_a}$.

Solution – E[Value] of a Record

- **`k`** is **`predictionKey(r)`**
- $\mathbf{s_a}$ is the *actual* class of **`r`**, rather than the observed class
- **`rso`** is the observed class of the record after translation from $\mathbf{S_r}$ to $\mathbf{S_l}$ and accounting for possible error via the discrimination matrix $\mathbf{DISCRIMINATOR_r}$

Solution – Evidence - Joint Sets

The BBN can incorporate evidence from objects in a *joint set*.

The joint set, $\mathbf{joint(R_1,R_2)}$ between any two record sets $\mathbf{R_1}$ and $\mathbf{R_2}$ is the set of all objects with a global key that is in both $\mathbf{R_1}$ and $\mathbf{R_2}$

For a record r that is known by AI, we can instantiate the **`lso`** node to add evidence to the translation.

Solution – Evidence - Experiments

Additional evidence can be accumulated by taking action on a translated record.

The observed outcome of the experiment is instantiated on the **e** node of the BBN.

Solution – Class Compression

Under certain conditions, we can compress the space of the translation and reduce computational efforts by pruning.

- Let $\mathbf{s}_1$ and $\mathbf{s}_2$ be elements of $\mathbf{S}_1$.
- $\mathbf{s}_1$ subsumes $\mathbf{s}_2$ if...

  $\mathtt{action}(\mathbf{A}_1, \mathbf{s}_1, \mathbf{s}_2) \geq \mathtt{action}(\mathbf{A}_1, \mathbf{s}_1, \mathbf{s}_1)$
- Let $\mathbf{s}_1' = \mathbf{s}_1 \cup \mathbf{s}_2$

Solution – Class Compression

Subsumption is transitive, allowing a hierarchy to be constructed over $\mathtt{s}_1$ by bottom-up association.

The compression of classes by subsumption is a function of an agent's perceptions, making the constructed hierarchy independent of any remote entities.

Solution – Class Compression

We introduce the operations...

- **`basis`**$(\mathbf{A_l})$ – The classes of $\mathbf{S_l}$ that cannot be subsumed.
- **`next`**$(\mathbf{S_{ls}}, \mathbf{s_l})$ – The immediate successors of $\mathbf{s_l}$ in the subsumption $\mathbf{S_{ls}}$ hierarchy of $\mathbf{A_l}$.

Solution – Algorithm – t value

Using what has so far been developed we can construct a function to determine the expected utility gain of applying translator **`t`** to remote data set.

Let the function $\mathbf{tvalue(t,R_r,A_l,S_{ls})}$ be defined as follows...

Solution – Algorithm – t value

```
value ← 0
for r in R_r:
  k ← predictionKey(r)
  rso ← class(r)
  rsot ← t(rso)
  d ← BBN(A_l,k,rsot)
  for s_l in S_ls:
     value ←value + E[action(A_l,rsot,s_l)]
```

Solution – Algorithm – t value

Where...

- **`t`** is a translator
- $\mathbf{R_r}$ is the remote record set
- $\mathbf{A_l}$ is the local agent
- $\mathbf{S_{ls}}$ is some compression (including no compression) of $\mathbf{S_l}$.

Solution – Algorithm

```
t ← default(Sr)
maxv ← tvalue(t,Rr,Al,Sl)
for sr in Sr:
  q ← basis(Al)
  while q not empty:
      sl ←pop(q)
      t′ ←update(t,sr,sl)
      tv ←tvalue(t,Rr,Al,Sls)
```

$$
\begin{aligned}
&t \leftarrow \mathrm{default}(Sr)\\
&maxv \leftarrow \mathrm{tvalue}(t, R_r, A_l, S_l)\\
&\mathbf{for}\ s_r\ \mathbf{in}\ S_r:\\
&\quad q \leftarrow \mathrm{basis}(A_l)\\
&\quad \mathbf{while}\ q\ \mathbf{not\ empty}:\\
&\qquad s_l \leftarrow \mathrm{pop}(q)\\
&\qquad t' \leftarrow \mathrm{update}(t, s_r, s_l)\\
&\qquad tv \leftarrow \mathrm{tvalue}(t, R_r, A_l, S_{ls})
\end{aligned}
$$

Solution - Algorithm

```
if tv > maxvalue:
        maxvalue ← tv
        t ← t′
        q ← next(S_ls, s_l)
```

At the end of the algorithm, **`t`** holds the best possible translator of the set $\mathtt{T}_{s_l s_r}$

Future work – Better Evidence

Currently, the evidence gathered by experimentation is underutilized.

Evidence is only propagated to a single record when it should enhance confidence in the entire translation.

Future work – EOG Graphs

Once attribute resolution has been concluded, EOG relationships between objects should be found.

This extension will allow taxonomies/ontologies to be resolved for a vastly more complete method for knowledge sharing.

Future work – Compound Classes

An agent's actions naturally partition a schema into classes. These classes can be constructed along multiple attributes. Eg. An agent takes action a on <M,+> and b on <F,0>

Simultaneous resolution of multiple attributes is much more complex than single attribute resolution, yet much more useful.

Conclusions

- The decision theoretic method presented here complements the linguistic and textual methods used more commonly in other research efforts.
- By explicitly incorporating a set of strategies and payoffs, we allow an agent to act with high autonomy when such actions are risky.

Conclusions

- The method for computing a translation's expected value has been improved by accounting for joint sets and experimental evidence.
- Class compression can lead to pruning of the search space with no loss of optimality. The value of this improvement will be more noticeable when resolving compound attributes.