6 Bibliography


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


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


Appendix – Setup and calibration of similarity models

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

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

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

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

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

The type of denotation sets adopted, the \( \text{sim} \) function and the similarity functions \( \sigma_i \) are the similarity model for the matching algorithms, and the
threshold value is the *calibration* of the similarity model.

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

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

However, to avoid *overfitting* of the similarity model with respect to the training corpus, we suggest using *cross validation*, which is a process with the following major steps
1. The training corpus is divided in \(n\) parts.
2. Each similarity model is calibrated with data of \(n-1\) parts and tested with the remaining \(n\) part.
3. Steps 1 and 2 are repeated for each of the \(n\) parts.
4. The final performance of each similarity model is the average overall performance \((f)\) for each of the \(n\) parts.

We describe such an evaluation in (Leme et al. 2008b), where we were interested in evaluating the best similarity model for property matchings. In this experiment, we used data extracted from the gazetteers Alexandria Digital Library and Geonames, and data extracted from eBay and Amazon. The denotation sets for properties were the set of tokens \(T\) and the set \(IV\) of pairs of the form \((\text{instance}, \text{token})\). We considered both sets as multisets, denoted \(\overline{T}\) and \(\overline{IV}\), respectively, and as sets in the usual sense, denoted \(T\) and \(IV\). We adopted *max* as the *sim* function, and the *cosine* with \(TF/IDF\), the *contrast model function* and the *information theory measure* as similarity functions. For the contrast model function, we used as parameter values \(\alpha=1.0\), and \(\beta\) and \(\gamma\in[1.0,10.0]\).
threshold varied from 0.0 to 1.0, in steps of 0.1.

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

Table 22. Automatically obtained vocabulary matching from eBay into Amazon

<table>
<thead>
<tr>
<th>Similarity models</th>
<th>Calibration</th>
<th>Selected model</th>
</tr>
</thead>
<tbody>
<tr>
<td>σ_i</td>
<td>α</td>
<td>β,γ</td>
</tr>
<tr>
<td>contrast model</td>
<td>[1.0,10.0]</td>
<td>max()</td>
</tr>
<tr>
<td>contrast model</td>
<td>[1.0,10.0]</td>
<td>max()</td>
</tr>
<tr>
<td>contrast model</td>
<td>[1.0,10.0]</td>
<td>max()</td>
</tr>
<tr>
<td>contrast model</td>
<td>[1.0,10.0]</td>
<td>max()</td>
</tr>
<tr>
<td>information theory</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>information theory</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>information theory</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>information theory</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cosine with TF/IDF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cosine with TF/IDF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cosine with TF/IDF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>cosine with TF/IDF</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>