Matchmaking augmented with general knowledge

Thomas Kleemann

Institut für Informatik, Universität Koblenz-Landau, D-56070 Koblenz, Germany*
tomkl@uni-koblenz.de

Description Logics (DLs) are well suited for the representation of terminological knowledge and support efficient reasoning about these terminologies. In this use case a large set of classified advertisements (ads) has to be matched with queries, derived from individual readers search requests as well as management needs to track the amount of classified ads in certain areas of interest.

Queries and ads are represented by concept expressions C_q and C_{ad}. An ad is a match of a given query iff $C_{ad} \sqsubseteq C_q$ with respect to the domain knowledge represented in the TBox, containing concepts and concept inclusions of the description logic knowledge base.

For every advertisement a concept expression is generated from the highly abbreviated text. Only the most necessary information is printed to save space and expenses. In addition to these economic reasons the reader shall not be bothered by redundant information. As a consequence only those features of goods are mentioned in ads, that are not commonly assumed. Caused by this restricted text the retrieval of the matching was weak. The common knowledge has to be included in the process of matching i.e. the subsumption has to recognize general knowledge and expectations that are not mentioned in C_{ad}. This inclusion has to be compatible with exceptions from the general.

To integrate this general knowledge with the matching process the TBox $\mathcal{T}$ is split into the standard part $\mathcal{T}_S$ and the general or default part $\mathcal{T}_D$. The general knowledge has to be applied, iff the subsumption is *uncertain* wrt. $\mathcal{T}_S$. A subsumption $C \sqsubseteq D$ is called uncertain $C \overline{\not\sqsubseteq}_{\mathcal{T}_S} D$ with respect to a TBox $\mathcal{T}_S$, if neither $C \sqsubseteq D$ nor $C \sqsubseteq \overline{D}$ are a consequence of $\mathcal{T}_S$.

Based on the split TBox and uncertainty the *default subsumption* can be defined. The default subsumption $C \sqsubseteq_{\mathcal{T}} D$ shall hold iff it is a consequence of $\mathcal{T}_S$ or - in case of uncertainty - a consequence of $\mathcal{T}_S \cup \mathcal{T}_D$ for a satisfiable concept C.

$$\mathcal{T} \vdash C \sqsubseteq D \iff \mathcal{T}_S \vdash C \sqsubseteq D \\ \vee ((C \overline{\not\sqsubseteq}_{\mathcal{T}_S} D) \wedge (\mathcal{T}_S \cup \mathcal{T}_D \vdash C \sqsubseteq D) \wedge (\mathcal{T}_S \cup \mathcal{T}_D \nvdash C \sqsubseteq \bot))$$

According to this definition the uncertainty persists, as long as possible models of the TBox exist, that support either $C \sqsubseteq D$ or $C \sqsubseteq \overline{D}$. The check for default subsumption will possibly constrain models in a way that one of the subsumptions holds.

This approach[1] to integrate common knowledge with matchmaking is compatible with DLs featuring the complement. Decidability is preserved if the extended DL is decidable. RACER's support for multiple TBoxes makes the implementation a breeze. The development of a specialized tableau is not needed. Well-known problems like the penguin or the platypus can be handled.

Unlike the standard subsumption the *default subsumption* revealed the matches, that a human reader of the ads marked positive.

* This research is funded by Stiftung Rheinland-Pfalz für Innovation.

[1] For further details and related work please visit http://www.uni-koblenz.de/~tomkl.