Learning Similarity Measures: A Formal View based on a Generalized CBR Model

Armin Stahl

Problem:

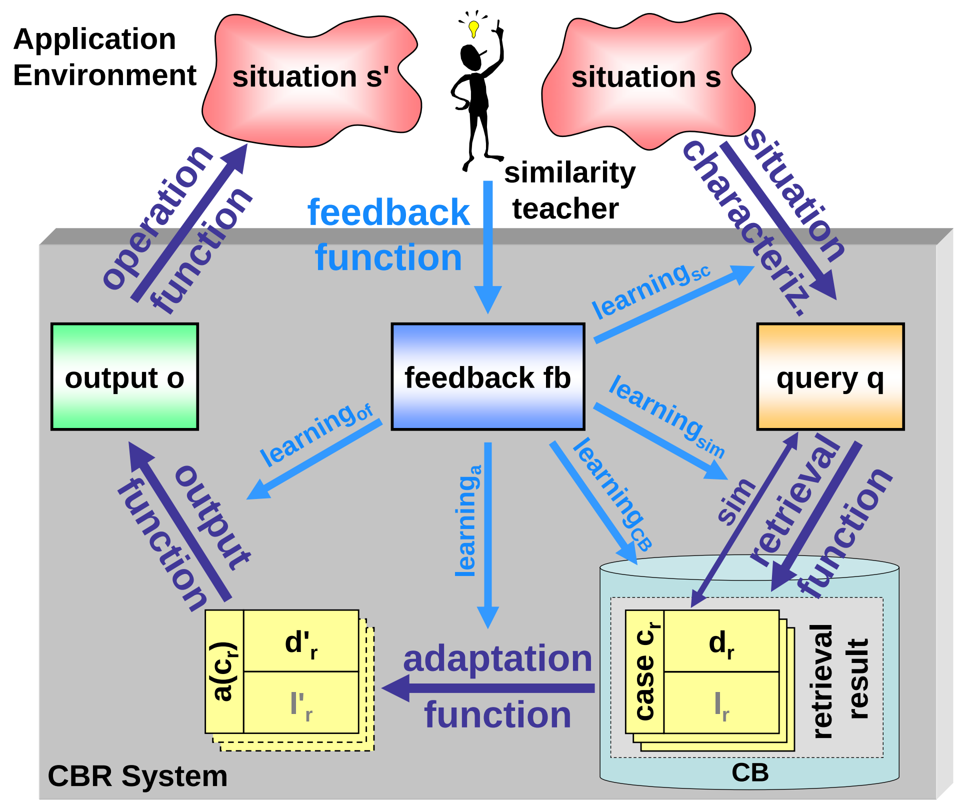
- Various approaches for learning similarity measures exist, but
- a clear methodology for applying them is still missing.
- **Important questions:**
  - What is the desired semantics of the similarity measure?
  - What kind of training data is suitable and how can it be acquired?
  - What type of similarity model has to be learned?
  - Which learning techniques are suitable to achieve best results?

Objective:

- Development of a formal foundation for
  - analyzing requirements
  - selecting accurate similarity models
  - acquiring training data
  - choosing accurate learning techniques
  - future research

A Formal and Generalized CBR Model:

- **Drawbacks of traditional CBR cycle:** [Aamodt & Plaza, 1994]
  - limited to pure problem-solving tasks
  - assumes cases to be problem-solution pairs
  - does not explicitly consider learning of general knowledge
  - neglects the interaction with application environment
  - does not consider novel CBR techniques (e.g. explanations)
- **Contributions of novel model:**
  - formal mathematical view
  - suited to describe arbitrary CBR applications
  - cases may represent arbitrary knowledge chunks
  - new phases for interaction with application domain (*sc*, *f*, *op*)
  - explicit processes for learning each knowledge container
- **Goal of CBR system:**
  Maximization of utility $u(s,o)$ of output $o \in O$ given situation $s \in S$ with respect to the domain specific utility function $u : S \times O \rightarrow [0,1]$



Goal of Learning Process:

- Acquisition of (partial) knowledge about domain specific utility function u
- Generation of measure sim which approximates $u(s,o_r)$ for all $s \in S$ and $c_r \in CB$

$$u(s,o_r) = u(s, of(q, a(q,c_r))) = u(s, of(sc(s), a(sc(s), c_r)))$$

Desired Semantics of Similarity Measures:

- **Determining the most useful case:**
  $$\arg\max_{c_r \in CB} sim(q,c_r) = \arg\max_{c_r \in CB} u(s,o_r)$$
- **Separating useful and useless cases:**
  $$\forall c_i \in CB^+, c_j \in CB^- : sim(q,c_i) > sim(q,c_j)$$
- **Ranking the most useful cases:**
  $$\forall c_i, c_j \in CB^u, c \in CB \setminus CB^u : \begin{array}{l} sim(q,c_i) > sim(q,c_j) \Leftrightarrow u(s,o_i) > u(s,o_j) \\ \wedge\, sim(q,c_j) > sim(q,c) \end{array}$$
- **Approximating the absolute utility of cases:**
  $$\forall c_i \in CB : sim(q,c_r) \approx u(s,o_r)$$

Training Data:

- **Absolute Case Utility Feedback (ACUF):**
  - training data provides absolute utility values, i.e. $u(s,o_j) = x$ with $x \in [0,1]$
- **Relative Case Utility Feedback (RCUF):**
  - the utility of outputs is determined relatively to other outputs, e.g. $u(s,o_i) > u(s,o_j)$

Future Work:

- Development of novel learning techniques for
  - approximating absolute utility values (→reliability)
  - extending probabilistic approaches to using RCUF
- Implementation of feedback-function f using HCI methods

Learning Techniques:

- **Gradient Search:**
  - suited for learning feature weights
  - can use ACUF and RCUF [e.g. Stahl, 2004]
- **Genetic Algorithms:**
  - suited for learning local similarity measures
  - can use ACUF and RCUF [Stahl & Gabel, 2003]
- **Probabilistic Similarity Models:**
  - known approaches only rely on ACUF [e.g. Breuel, 2003]

IUPR

German Research Center for Artificial Intelligence (DFKI)
Image Understanding and Pattern Recognition Group
Kaiserslautern, Germany

Contact:
www.iupr.org/~stahl
Armin.Stahl@dfki.de