Revisiting Associative Learning

John Anderson
Dan Bothell ?
Christian Lebiere?
Carnegie Mellon University

Rational Analysis: Bayesian Formula for Activation of Chunks in Declarative Memory

- Activation of a memory should scale with the log odds that the memory is needed.
- The Bayesian formula for the posterior odds of needing a memory i in a context:

$$\frac{P(N_i|Context)}{P(\sim N_i|Context)} = \frac{P(N_i)}{P(\sim N_i)} \times \frac{P(Context|N_i)}{P(Context|\sim N_i)}$$

- A_i=log(posterior odds)

  = log($\frac{P(N_i)}{P(\sim N_i)}$) +$\sum_{C_j \epsilon Context} log(\left[\frac{P(C_j|N_i)}{P(C_j|\sim N_i)}\right]$

  = $B_i + \sum_j W_j S_{ji}$

- The question of interest: How are the S_{ji}'s learned?

Rational Analysis: "Bayesian" Formula for S_{ji}

- S_{ji} are calculated from a weighted average of a prior estimate, R_{ji}, and an empirical estimate, E_{ji}, of the likelihood ratio

$$\frac{P(C_j|N_i)}{P(C_j|\sim N_i)} \approx \frac{P(N_i|C_j)}{P(N_i)}$$

- $S_{ji} = log\left(\frac{assoc \times R_{ji} + F(C_j) \times E_{ji}}{assoc + F(C_j)}\right)$ where *assoc* is a parameter and $F(C_j)$ is the frequency of cue j.
- R_{ji}=$\frac{1/n}{1/m}$=m/n for connected chunks (1 otherwise) where m is all chunks and n is number of chunks connected to cue j.
- This leads to the non-learning formula S_{ji}= S –log(n) for connected chunks and 0 otherwise.
- $E_{ji} = \frac{F(N_i \& C_j)/F(C_j)}{F(N_i)/F}$ where F(X) denotes the frequency of X.

Some of the Problems with Formulation

- **All experiences represented equally**: Experimental effects like the fan effect would be drowned out if the huge prior set of associations were represented.
- **Awkward combination of prior and empirical:** The first time a j-i combination is experienced, there can be an abrupt shift from a prior of 0 to a large negative value.
- **Enormous Storage Demands**: Grows with the square of number of chunks although many combinations seem superfluous (e.g., 2+6=8 chunk priming a visual chunk).
- **Irrelevant cues:** Elements in a buffer that are there for irrelevant purposes (part of computation) get counted as cues both for creating associations and as retrieval cues.
- **Role-Independent Associations:** For instance 2 and 6 in the query 2+6=? spread as much activation to 2+4=6 as to the desired 2+6=8.

VISCA Talk: Use of S_{ji} for Causal Attribution

		Effect	
		Yes	No
Cause	Yes	a	b
	No	c	d

a, b, c and d are frequencies with which subjects observed the presence and absence of a possible cause with the presence or absence of an effect.

- ACT-R model attributed a cause-effect relationship when a pairing became reliable enough for the effect to be retrieved.
- sji-hook for strength of association S_{ji} from cause (j) to effect (i):
  $S_{ji} = \log\left(\frac{P(j|i)}{P(j|{\sim}i)}\right) = \log\left(\frac{a/(a+c)}{b/(b+d)}\right)$=log(a)-log(a+c)-log(b)+log(b+d)
- The model like Cheng's power-PC correctly predicts that subjects are largely insensitive to sample size but rather just relative sizes of a, b, c and d and that they weigh a and b more.
- Could this be used as a template for a general associative learning approach between cues (causes) and chunks (effects).
- Note the a, b, c and d in the the causal implementation where 1 plus the actual frequencies – a better way to incorporate priors.

Fan Effect Analysis

		Memory Chunks			
		Doctor-Bank	**Hippie-Park**	Hippie-Church	Others
Cues	**Doctor**	10	0	0	10
	Bank	10	0	0	10
	Hippie	0	10	10	10
	Park	0	10	0	10
	Church	0	0	10	10
	Others	0	0	0	1000

The summed weights of all the Others could only be this weak because of decay.

+1 prior

$$S_{ji} = \log\left[\frac{F(N_i \& C_j)/F(C_j)}{F(N_i)/F}\right]$$

		Doctor-Bank	
		Needed	Not Needed
Doctor	Present	11	11
	Absent	11	1081

F(Ni&Cj)/F(Cj)=11/22=.50

F(Ni)/F=11/1114=.01

S_{ji}=1.65

		Hippie-Park	
		Needed	Not Needed
Hippie	Present	11	21
	Absent	11	1071

F(Ni&Cj)/F(Cj)= 11/32=.37

F(Ni)/F=11/1114=.01

S_{ji}=1.35

Implementation Details

$$B_i + \sum_j W_j S_{ji}$$

1. When counts and S_{ji}s are updated: Clearing of chunk from Imaginal or entry of chunk into Retrieval.
2. Cues (the j's) are the elements of the chunks in Imaginal.
3. The memories (i's) are only things created in Imaginal or retrieved.
4. The storage requirements are probably much less than the square of the number of all chunks.
5. Use the approximation: $S_{ji} = log\left[\frac{F(N_i \& C_j)/F(C_j)}{F(N_i)/F}\right]$
6. The frequencies are seeded with non-zero priors.
7. $F(X) = a + \sum_{k \in X updates} t_k^{-d}$ where t_k is the time since the kth update involving X.

Potential Predictions of Proposal

- Fan effect: Interfering effect of number of things learned about an element on retrieval of those things when cued with the element.
- This interference really reflects the decrease in the probability of a memory in the presence of the cue and not number of interfering associations.
- Interfering effect of number of things learned about an element on retrieval of pre-experimental knowledge about the element.
- With power-law decay their influence on pre-experimental memories decreases with passage of time.
- If one practices retrieving a paired memory (A-B) from one element A, it will become more accessible from element A than B.

Problems with the Old Formulation

- **All experiences represented equally**: Experimental effect like the fan effect would be drowned out if the huge prior set of associations were represented.
- **Awkward combination of prior and empirical:** The first time a j-i combination is experienced, there can be an abrupt shift from a prior of 0 to a large negative value.
- **Enormous Storage Demands**: Grows with the square of number of chunks although many combinations seem superfluous (e.g., 2+6=8 chunk priming a visual chunk).
- **Irrelevant cues:** Elements in a buffer that are there for irrelevant purposes (part of computation) get counted as cues both for creating associations and as retrieval cues.
- **Role-Independent Associations:** For instance 2 and 6 in the query 2+6=? spread as much activation to 2+4=6 as to the desired 2+6=8.