Experimental Evaluation of Similarity Measures for Educational Items

Jaroslav Čechák Radek Pelánek

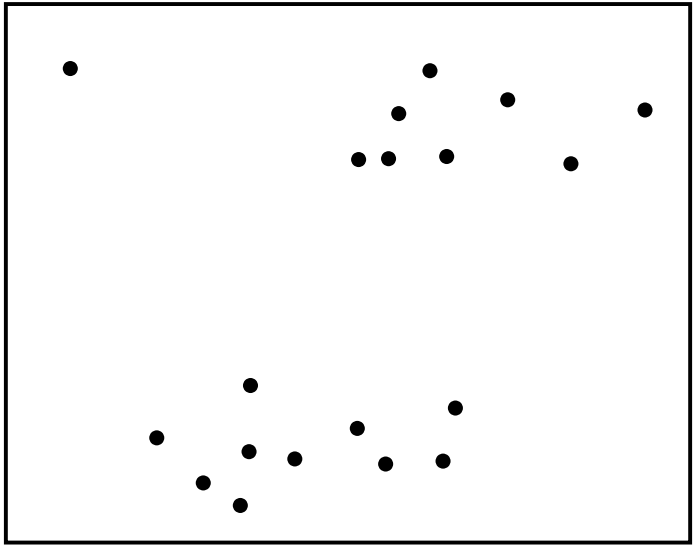
EDM 2021

Problem

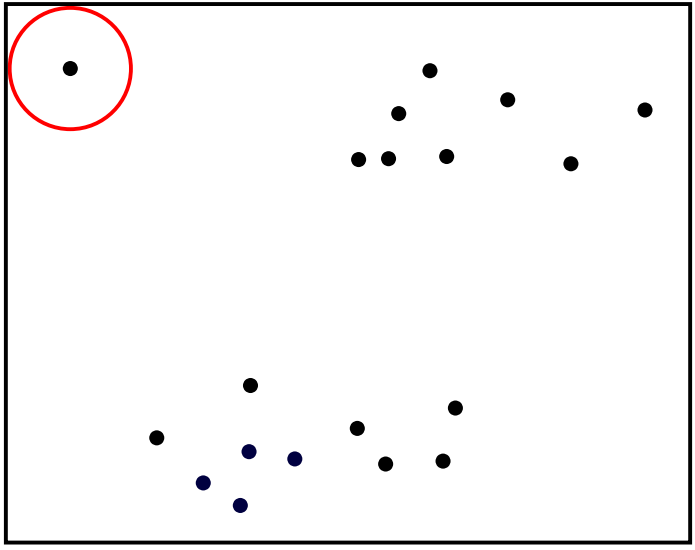
item stem	correct	distractor	explanation
I _ to the gym once a week.	go	am going	When talking about periodical events, we use present simple tense.
He _ breakfast every day.	eats	is eating	When talking about periodical events, we use present simple tense.
I _ this book.	like	am liking	When talking about general state, we use present simple tense. The verb to like is not used in continuous form.

54 knowledge components, 68 item sets, 4 348 items

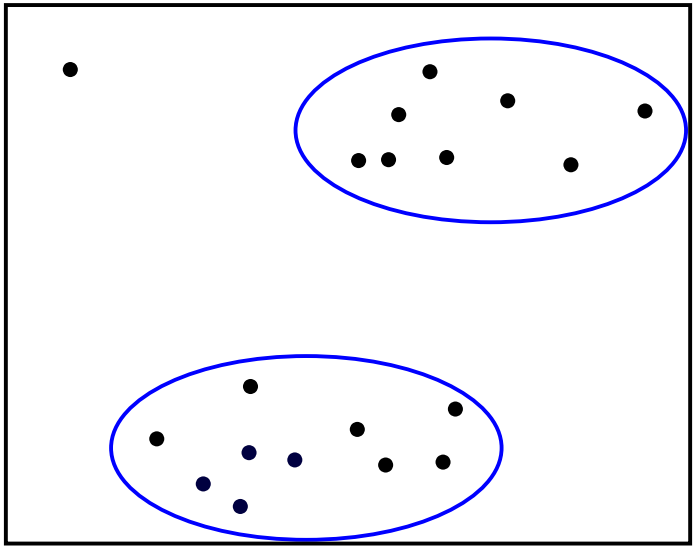
Solution



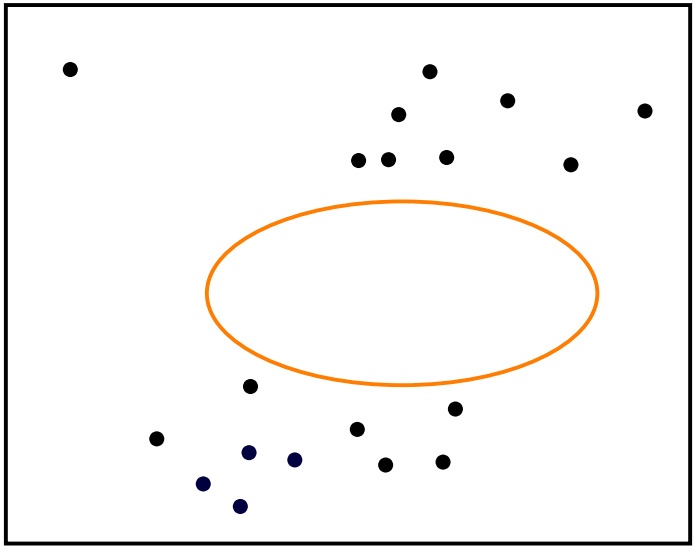
Solution



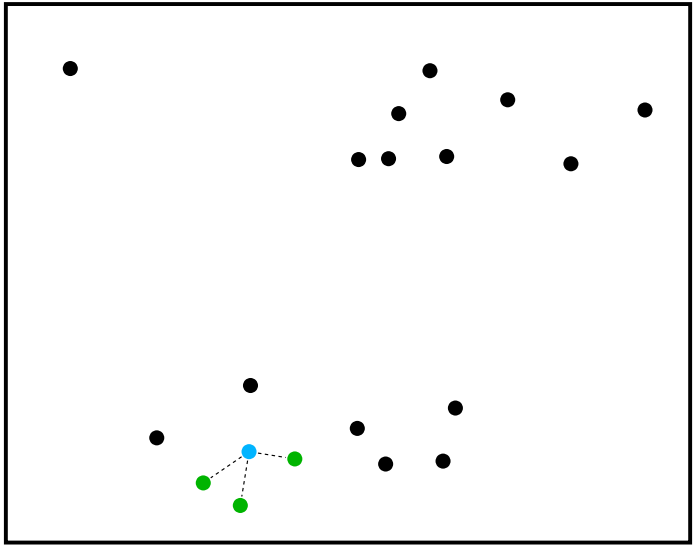
Solution



Solution



Solution



Data

item stem	correct	distractor	explanation
I _ to the gym once a week.	go	am going	When talking about periodical events, we use present simple tense.

student	item	correct	response time
123	10	1	5500

Explanation similarity

E1: When talking about periodical events, we use present simple tense.

E2: When talking about general state, we use present simple tense. The verb to like is not used in continuous form.

Jaccard Index: $\frac{|E_1 \cap E_2|}{|E_1 \cup E_2|} = \frac{8}{8+14} = 0.36$

Performance similarity

	items	
student	i	j
1	1	1
2	0	1
3	1	0
4	0	0
5	1	0
6	1	1
7	0	1

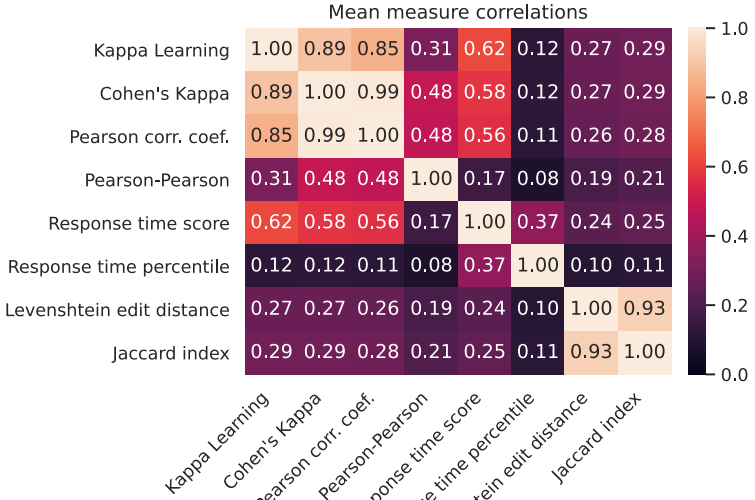
$$S_p = \frac{(ad - bc)}{\sqrt{(a+c)(a+b)(b+d)(c+d)}}$$
$$= \frac{(2 \cdot 1 - 2 \cdot 2)}{\sqrt{(2+2)(2+2)(2+1)(2+1)}}$$
$$= -0.25$$

$S_p \equiv$ Pearson corr. coef.

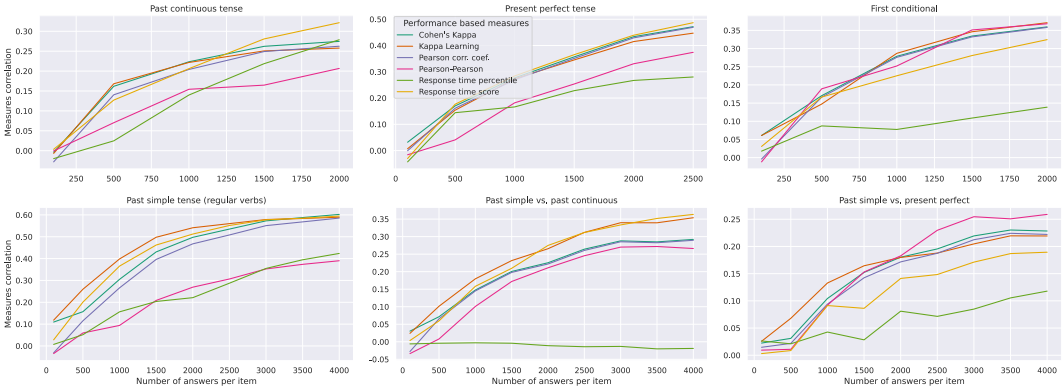
Similarity measures

name	measure type	data used
Levenshtein edit distance	content	explanations
Jaccard index	content	explanations
Pearson corr. coef.	performance	correctness
Cohen's Kappa	performance	correctness
Kappa Learning	performance	correctness
Pearson-Pearson	performance	correctness
Response time percentile	performance	response time
Response time score	performance	correctness + response time

Relations among measures



Amount of data required



Difference among knowledge components

- Correlations range from 0.06 to 0.67.
- The choice of knowledge component is more significant than the choice of measures.
  - Rule-based components have higher correlations on average.
  - Items for lower grades have higher correlations.
  - Newly added items lower the correlations.

Message

1. Performance-based similarity measures are data-hungry.

2. Results from one knowledge component may not generalize to others.

Thank you for your attention.

Experimental Evaluation of Similarity Measures for Educational Items
J. Čechák, R. Pelánek