mpii
max planck institut
informatik

Transitivity, Time Consumption, and Quality of Preference Judgments in Crowdsourcing

Kai Hui, Klaus Berberich

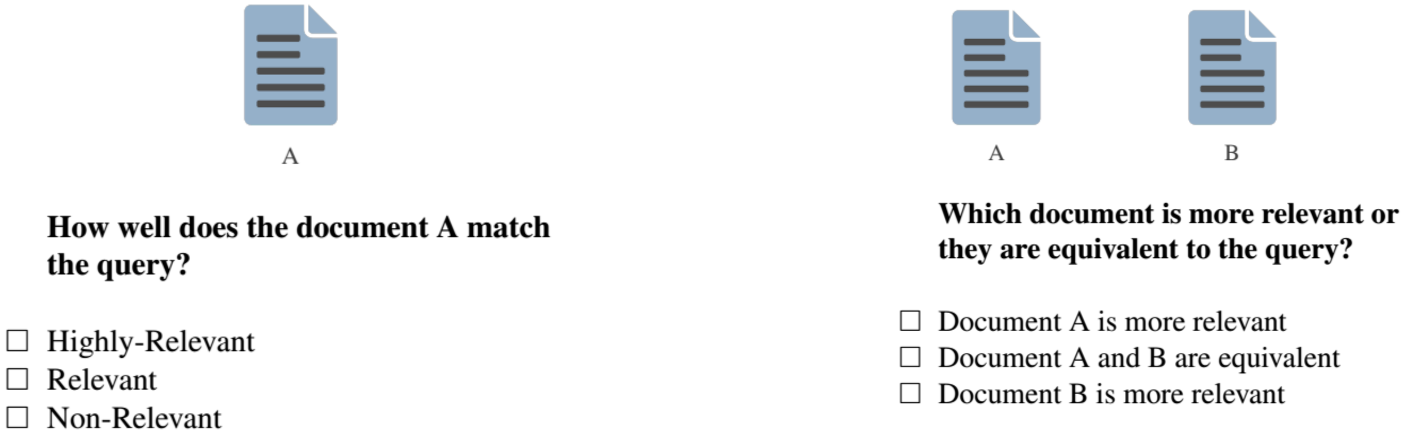
Max Planck Institute for Informatics

{khui, kberberi}@mpi-inf.mpg.de

ECIR
2017

Background

- There exist two kinds of manual judgments: graded judgments and preference judgments.



How well does the document A match the query?

☐ Highly-Relevant
☐ Relevant
☐ Non-Relevant

Which document is more relevant or they are equivalent to the query?

☐ Document A is more relevant
☐ Document A and B are equivalent
☐ Document B is more relevant

Background

- ❑ Preference judgments have been demonstrated to be a better alternative, but are very expensive:

  $O({N_d}^2)$ for N_d documents, and $O(N_d log N_d)$ when assuming transitivity.

- ❑ Strict and weak preference judgments are both widely employed in the literature

  Strict Preferences: $d_1 \prec d_2 \prec d_3 \prec d_4 \prec d_5$

  Weak Preferences: $d_1 \prec d_2 \sim d_3 \sim d_4 \prec d_5$

- ❑ Crowdsourcing provides a cheaper option

Research Questions

- ❑ Do weak/strict preference judgments exhibit transitivity when collected using crowdsourcing?

  Transitivity is crucial in reducing the number of preference judgments.

- ❑ How do weak/strict preference judgments compare against graded judgments in terms of time consumption?

  Fewer time consumption means one could pay less for preference judgments.

- ❑ Can weak/strict preference judgments collected using crowdsourcing replace judgments by trained judges?

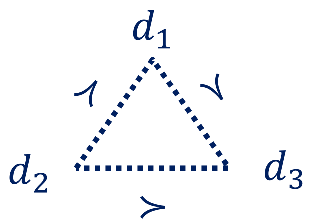
  Compare the quality of judgments from these three kinds.

Crowdsourcing

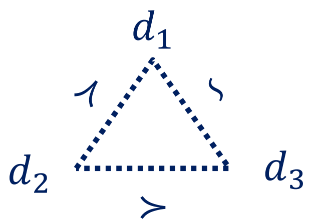
- Collect graded judgments, strict and weak preference judgments for twelve queries from TREC Web Track via CrowdFlower platform
- Measure time consumption when CrowdFlower workers make judgments
- Compare collected judgements in terms of their agreements to the judgments from TREC

Transitivity

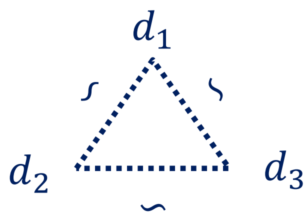
For document triples, count the ones which are transitive.



asymTran



s2aTran



s2sTran

Type of Preference Judgements		# Transitive Triples / # Total	Average Percentage
Strict Preferences	asymTran	212/220	96%
Weak Preferences	asymTran	46/47	98%
	s2aTran	98/108	90%
	s2sTran	21/65	32%
	Overall	164/220	75%

- □ Transitivity holds among strict preferences
- □ Transitivity does not hold among tie judgments

Time Consumption

Time Consumption (s)		Average	25th	Median	75th
Graded Judgments	# Judgment	2,60	1,37	1,52	1,82
	# Total	24,24	11,73	19,55	28,88
Strict Preferences	# Judgment	1,79	1,24	1,37	1,58
	# Total	34,17	17,84	25,28	40,98
Weak Preferences	# Judgment	2,07	1,40	1,57	1,91
	# Total	32,43	15,77	54,57	39,10

- Judges are faster in making strict preference judgments
- When considering total time (judgment time + reading time), judges need more time in preference judgments

Judgment Quality

Type of Judgement	Percentage Agreement	Cohen's κ
Graded Judgements	53%	0,282
Strict Preferences	74%	0,530
Weak Preferences	61%	0,419

- Judgment quality in terms of agreements relative to TREC judgments
- Preference judgments lead to significantly better quality
- Strict preference judgments are significantly better than weak preferences

Thank You!