Employing User Relevance Assessments for Measuring Retrieval Effectiveness

David J. Harper
School of Computing
The Robert Gordon University

Outline of Talk

- Gedanken experiment
- TREC8 interactive track
- Issues arising when using TREC8 interactive track collection
  - Recall depth
  - Judging relevance to topic
  - TREC assessor judgements
- Effect of topic interpretation on user studies
- Employing user relevance assessments for measuring effectiveness
- Discussion and conclusions

Gedanken experiment

Relevance Assessments for a Topic									
	Users								
	U1	U2	U3	U4	U5	U6	U7	Prop. #R	R/N?
Doc 1	N	R	R	R	R	N	R	5/7	
Doc 2	N	N	R	R	R	R	N	4/7	
Doc 3	R	R	N	N	N	N	R	3/7	
Doc 4	R	N	R	N	N	N	N	2/7	

TREC assessor

TREC8 Interactive Collection

- Collection: 210K FT documents
- Topics based on *ad hoc* topics, **without** narrative
- Aspectual retrieval task: save documents covering various aspects of topic
- Topic 408i: "tropical storms that cause loss of life or damage", find as many different storms as possible
- Highly successful in provoking intensive interaction by users

TREC8 Topic 408i

Number: 408i; **Title**: tropical storms

Description: What tropical storms (hurricanes and typhoons) have caused property damage and/or loss of life?

Instances: In the time allotted, please find as many DIFFERENT storms of the sort described above as you can. Please save at least one document for EACH such DIFFERENT storm. If one document discusses several such storms, then you need not save other documents that repeat those, since your goal is to identify as many DIFFERENT storms of the sort described above as possible.

Issues (1): Recall Depth

- Pooled results from 7 participating groups
- 6 topics, 1189 relevance assessments
- Aspectual relevance assessments: assessors had to assess which aspects (of many) each document addressed
- Likely that new interactive studies will lead to retrieval of **unassessed** documents
- In a recent study, among 415 unique documents saved by the users, 119 documents had no assessment in the QRELS (n.b. only 4 topics used of 6)

Issues (2) – Judging relevance

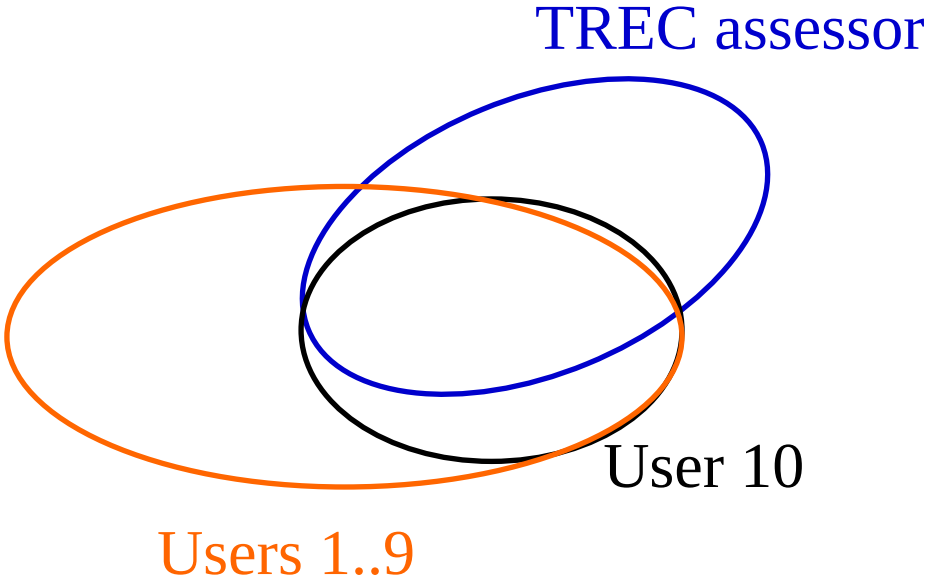
- Study authors (Kelly, Harper) judged relevance of 119 documents independently
- Level of agreement: 428i (91%), 438i (83%), 431i (100%), 408i (48%)
- Problematic topic 408i: differences revolved around interpretation of topic: what was meant by “damage”, “property”, “different storm”
- Similar remarks apply to other topics even given good levels of agreement above

Issues(3) – TREC8 assessor judgements

- Assessors judged relevance based on the topic as given, i.e. no narrative
- Assessors clearly had to settle on a particular interpretation of the topic, e.g. what they understood "damage" etc to mean in topic 408i
- TREC assessors judgements are considered as the "gold standard" but why?
- Why are the interpretations of the users participating in a study any less valid?

Effect of interpretation on measuring effectiveness

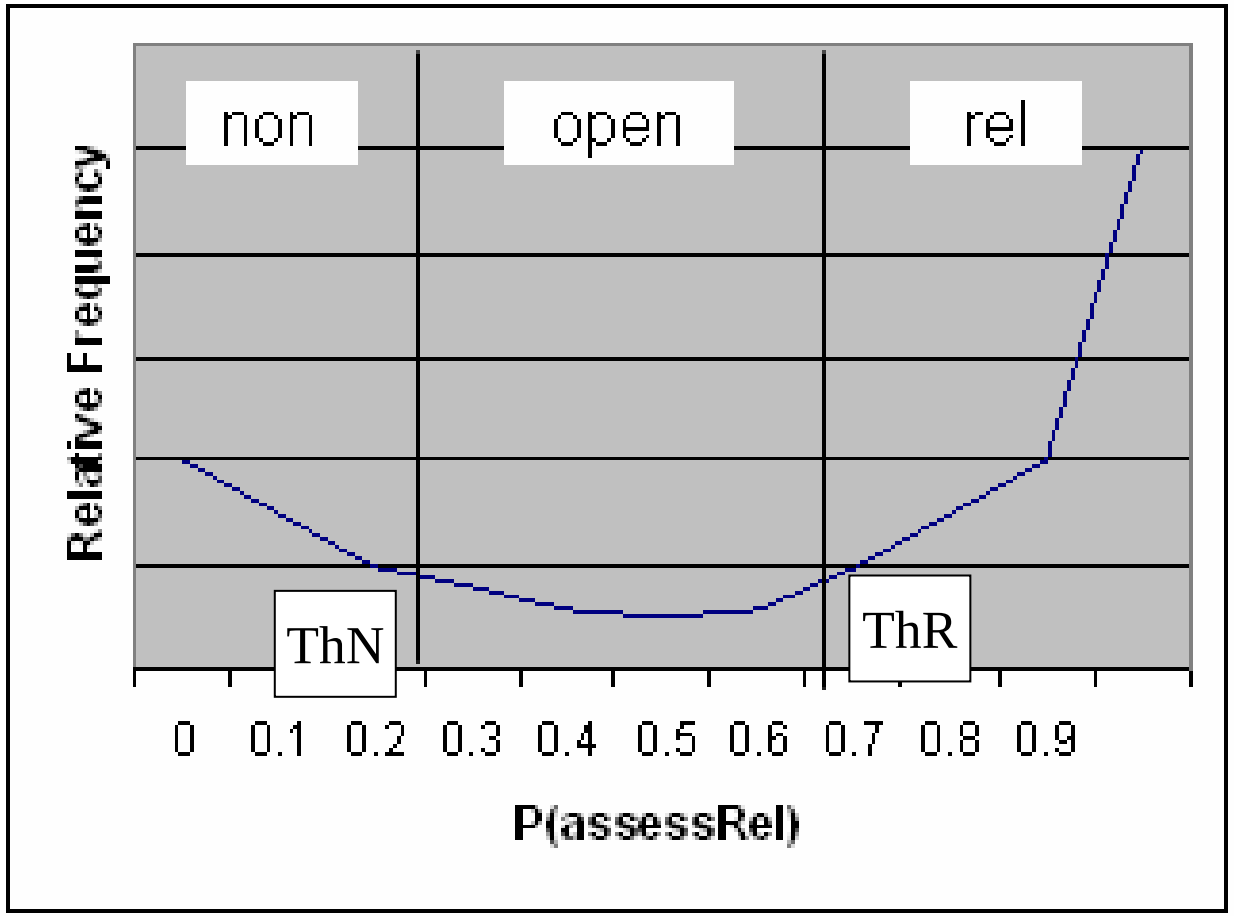
- Performance of users with differing interpretations lower
- Precision may simply be a measure of agreement between user and assessor – typical P values 0.6..0.8
- Differences between systems under study may be masked by differences due to differing interpretations of users.



Using user assessments for measuring effectiveness

- Typically, in a user study, users will save documents (assume these **relevant**), and display and not save others (assume **not relevant**)
- For a given topic, and for each document saved by at least one user:
  - # users who saved document (assessed relevant)
  - # users who viewed but did not save (assessed non)
  - Compute *P(doc assessed rel), P(assessRel)*
- Consider values *P(assessRel)* can take:
  - High, near 1.0, document likely relevant
  - Low, near 0.0, document likely not relevant
  - "In between", around 0.5, document subject to differing interpretations of topic

Using distribution of P(assess) values over topic for saved docs



Questions and Implications

- Which set of users should we use to establish pool of relevance assessments?
  - Users participating in a given study?
  - User participating in a number of previous studies?
- Distribution plots of *P(assess)* could be used to:
  - Determine degree to which a topic admits of multiple interpretations
  - Effect of task on relevance assessment
  - Explore performance of individual users
- Effect of recall depth when using TREC8 collection
- Implications for the evaluation of operational systems?