

It's nice and warm in the cave



Why are we here?

- IIR ignored?
- SIGIR
 - Full of cave men (some women)
 - Who drive drag racers

Feel the need for speed

- Speed is critical to IIR

Test collections

- Set of
 - documents
 - topics
 - relevance judgements
- Corner stone of evaluation



Test collections are for

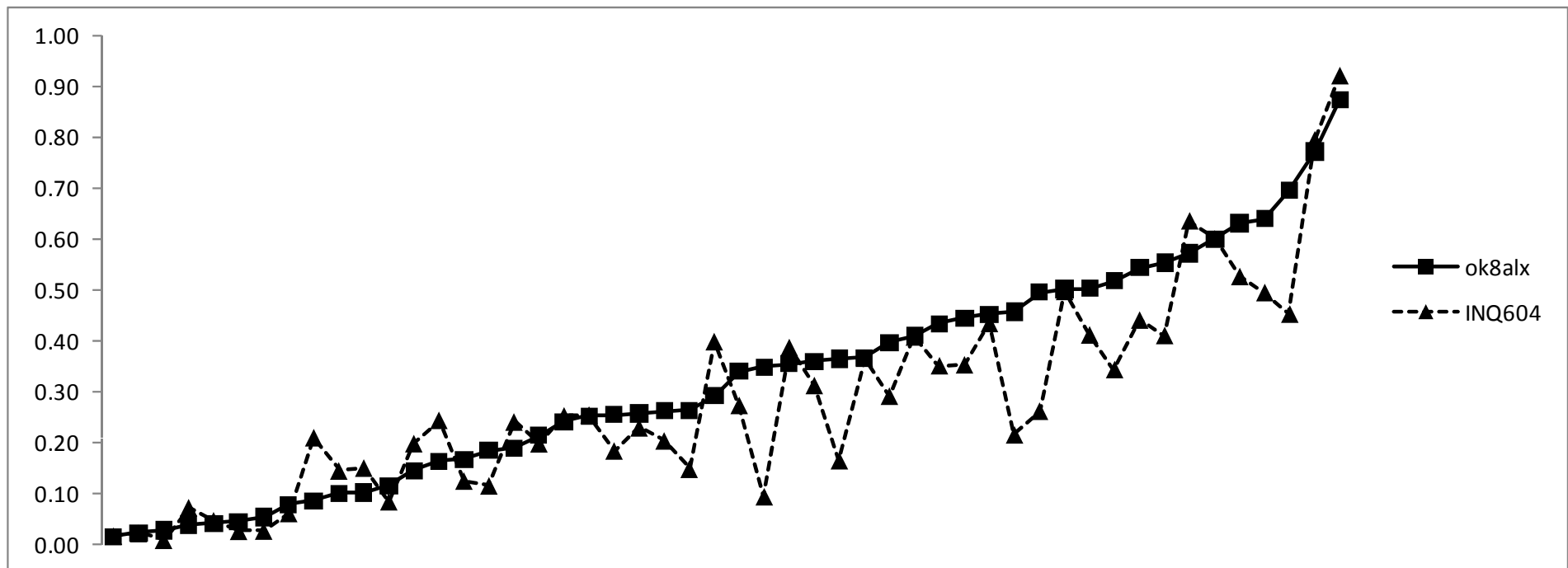
- Comparisons
 - Has my system retrieved more relevant documents than another?
- User model?
 - Users will ????? the system retrieving (on average) more relevant documents
 - No matter what

Problems?

- Turpin and Hersh, 2000, 2001
 - Two systems
 - Large test collection differences measured
 - 20 minutes searching
 - No user differences measured
- Bunch of other papers
 - Some supportive
 - Some opposing
 - Some either

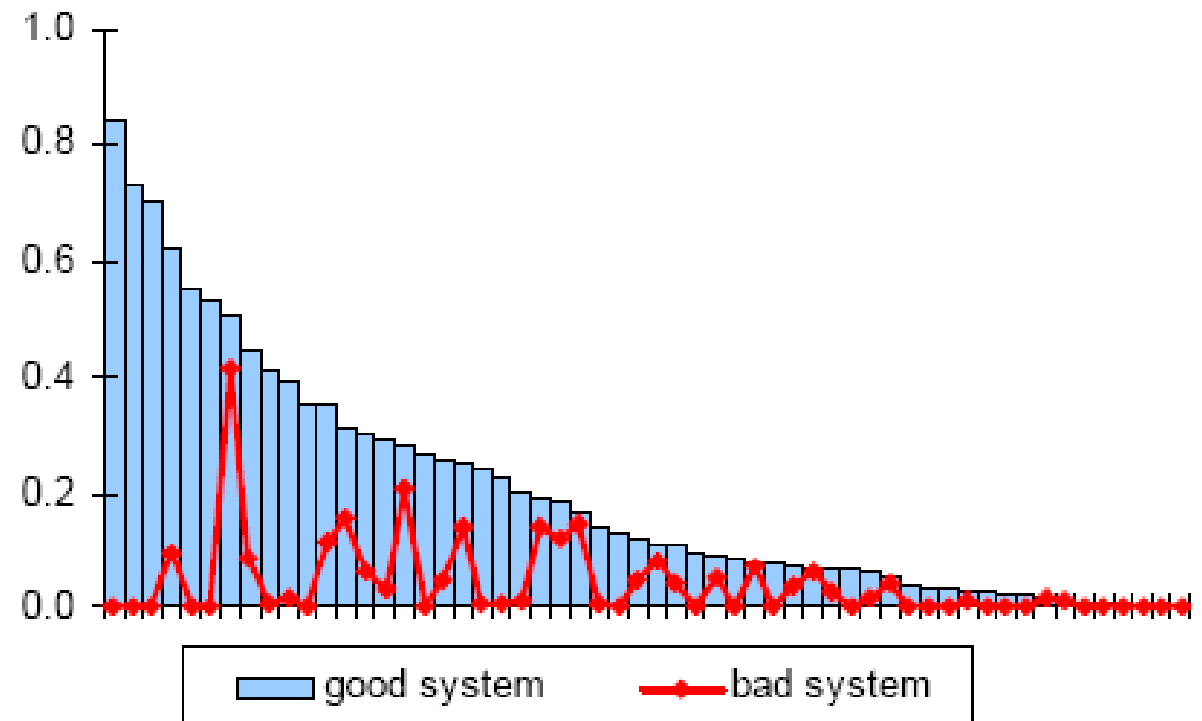
Topic variation

- ok8alx 0.324 – 15% relative improvement
- INQ604 0.281



Azzah Al-Maskari

- Two from three
- For each topic
 - Test three systems
 - Pick best
 - Pick worst
 - Observe users



Measured

- Time to find first relevant
 - TREC
 - Their judgements of relevance
- Number documents saved
- User satisfaction

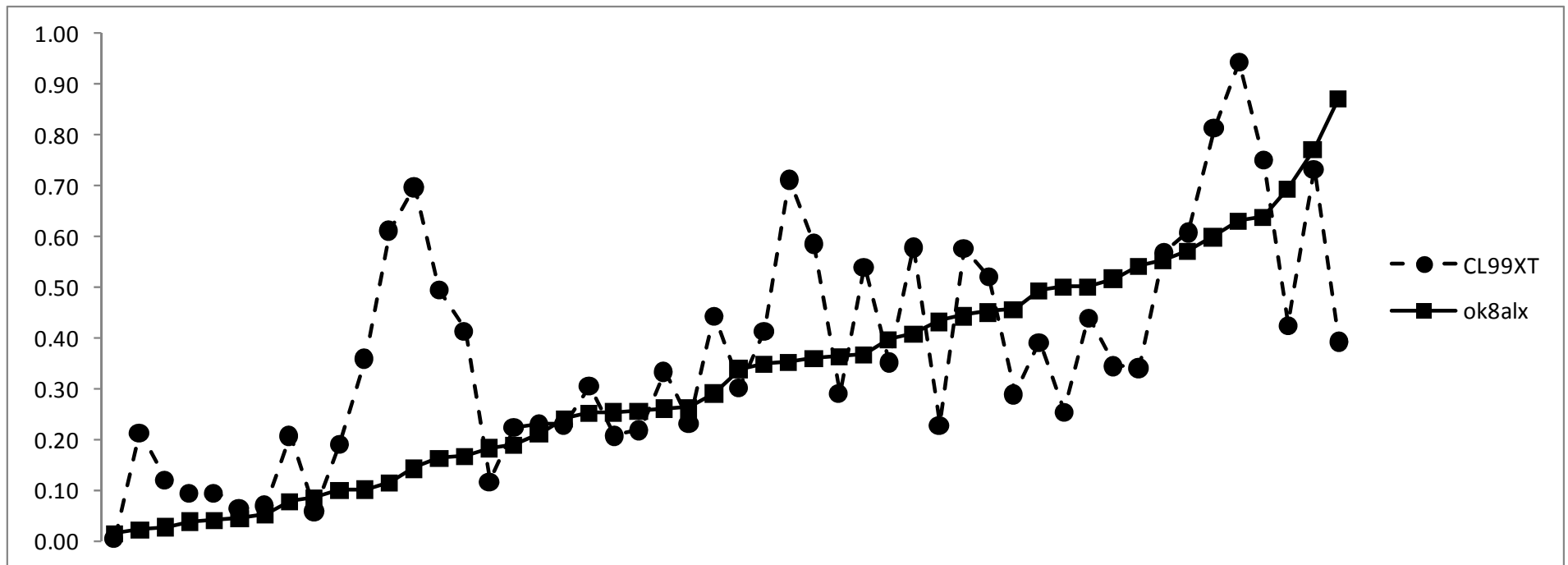
- Clear, significant differences

Smaller differences

- User measures harder to discern
- Like Thorsten's experiment
 - Give a marginally poorer system
 - Users cope may even make do
- Remember
 - ok8alx 0.324 – 15% relative improvement
 - INQ604 0.281

Topic variation

- CL99XT 0.373
- ok8alx 0.324



User experiments aren't usable

- Slow to set up
- Expensive
- Probably not big enough

- 50 topics?
 - 200?
 - 2,000?

Test collection problems

- Relevance feedback
 - Strong test collection evidence for it
- Automatic query expansion (pseudo-relevance feedback)
 - Strong test collection evidence for its inclusion
- Very poor take up of these technologies
 - Problem there

More problems

- Not enough of them
 - Obsession with TREC
- Lack of take up of technologies
 - Cross topic effects
 - Lack of control
- No capturing of interaction
 - Search output influencing subsequent actions
- Dependencies between relevant documents
- Upper bound on usefulness?
 - Can't think of a paper on this