Modern Day Search Engines

Review - Summary of Classic IR

- (add slide)

Divide between classical vs modern

- Thinking about the user, if they're satisfied, etc.

Are we really satisfying our users ?

1) can our previous measures really tell us user preferences ?
2) which one can better predict?
3) Will this power vary ? When choosing one system over anoter what are pros/cons

Experiment Settings

- we have users , systems -> check if they align in terms of their ranking/satisfaction
- where ? via crowd sourcing . **mechanical turk**
- test collection ? trec web track to evaluate ranking system against user eval & crowd sourcing
- When metric shows larger diference = Shows system is way better
- When if one system did not retrieve anything relevant ? . . .
- All systems returned relevant results ? Which metric is more predictive for users. interesting outlier p@k doesnt check about position of top k as long as it is in k. p@10 cant distinguish between systems

Conclusion

- Optimistic view. Classical metrics do decently
- effectivness of metrics vary
- Correlation is strong when performance different is large(TODO: review this)

User behavior oriented retrieval eval

- Cheap, large scale, natural usage context (no need to pretend or assume user behavior)
- Modern systems use A/B test

A/B Test

- Two-sample hypothesis testing

- Two versions a and b are comapred. which are identical except for one variation that might affect a users' behavior. I.E. indexing with/without stemming
- Randomized experiment ? Seperate population into equal size groups.
- Null hypothesis : no difference between system A and B -> z-test , t-test