Searcher's Feedback Quality and Effort in Interactive IR: A Simulation based on User Models

Kalervo Järvelin &
Heikki Keskustalo & Ari Pirkola
Department of Information Studies
University of Tampere, Finland

Outline

1. Why simulate interaction?
2. User models
3. Effects of quality and quantity of RFB
4. Conclusion & Future Work

Are system adaptation possibilities identified?

Why Simulate Interaction?

- Real interactive RFB tests:
  - experimental system + control system
  - 8, 16, 32 ... test searchers ($$$)
  - 4, 8, 16 test topics (hardly more)
  - Latin Square design to fight learning effects
  - e.g. 20 min per searcher & topic
  - rerun = new topics or test persons
  - provides nearly <u>real</u> and <u>rich</u> data
    - but not necessarily recommendations

Why Simulate Interaction?

- Simulated RFB tests:
  - any number of systems / modifications
  - easily any number of test searchers
  - easily any number of test topics
  - no learning effects
  - full test cycle = less than a day
  - rerun = just do it
  - data: if it only was real ...
    - may gain evidence for recommenda[…] on optimal adaptive behaviour

= what would happen if the users would behave as simulated? Can we advice them suitably?

User Models for RFB

- Simulation requires a user model that represents assumptions about <u>relevant aspects</u> of searcher behavior w.r.t RFB
- Obvious candidates:
  - relevance requirement by the searcher
  - value of relevance - tolerance of non-relevance
  - willingness to browse initial results
  - willingness to provide FB
  - level of topic understanding - consistency of FB

A Simple User Model Example

Model **M = <R, B, F>**

- Relevance threshold **R** to accept a document as FB document: **R** $\in$ {0, 1, 2, 3}
- Browsing window size: at most **B** top documents are browsed: **B** $\in$ {1, 5, 10, 30}
- Feedback set size: at most **F** feedback documents are collected: **F** $\in$ {1, 5, 10, 30}

All variables -> yield many *searcher scenarios*

Quality and Quantity of RFB

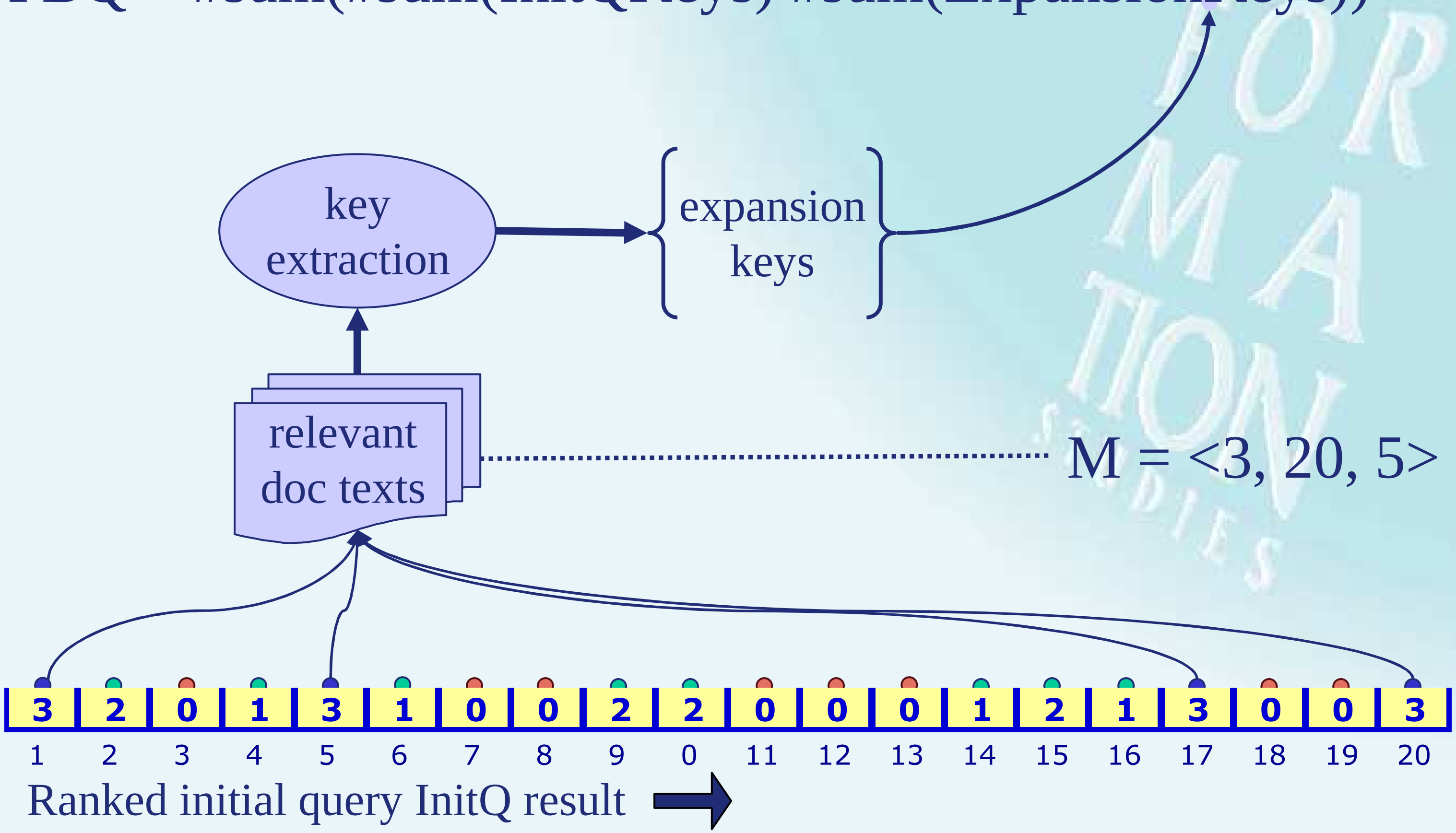
- ECIR'06
- Background:
  - Users might wish to find especially highly relevant documents (Kekäläinen & Järvelin)
  - Users are able to identify highly relevant documents (Sormunen & Vakkari)
  - In highly relevant documents ... (Sormunen & al.)
    - a larger share of aspects of the request topic is discussed
    - a larger set of unique expressions is used

Quality and Quantity of RFB

- Research questions:
  - How is the quality and quantity of RF related to search effectiveness?
  - How effective is RF when we consider relevance levels in evaluation?
  - How effective is RF compared to pseudo RF?

Basic Feedback Model Ex

FBQ = #sum(#sum(InitQKeys) #sum(ExpansionKeys))



Liberal Evaluation (MAP (%), N=41, TREC). Baseline MAP=20.7%

Good RFB not competitive

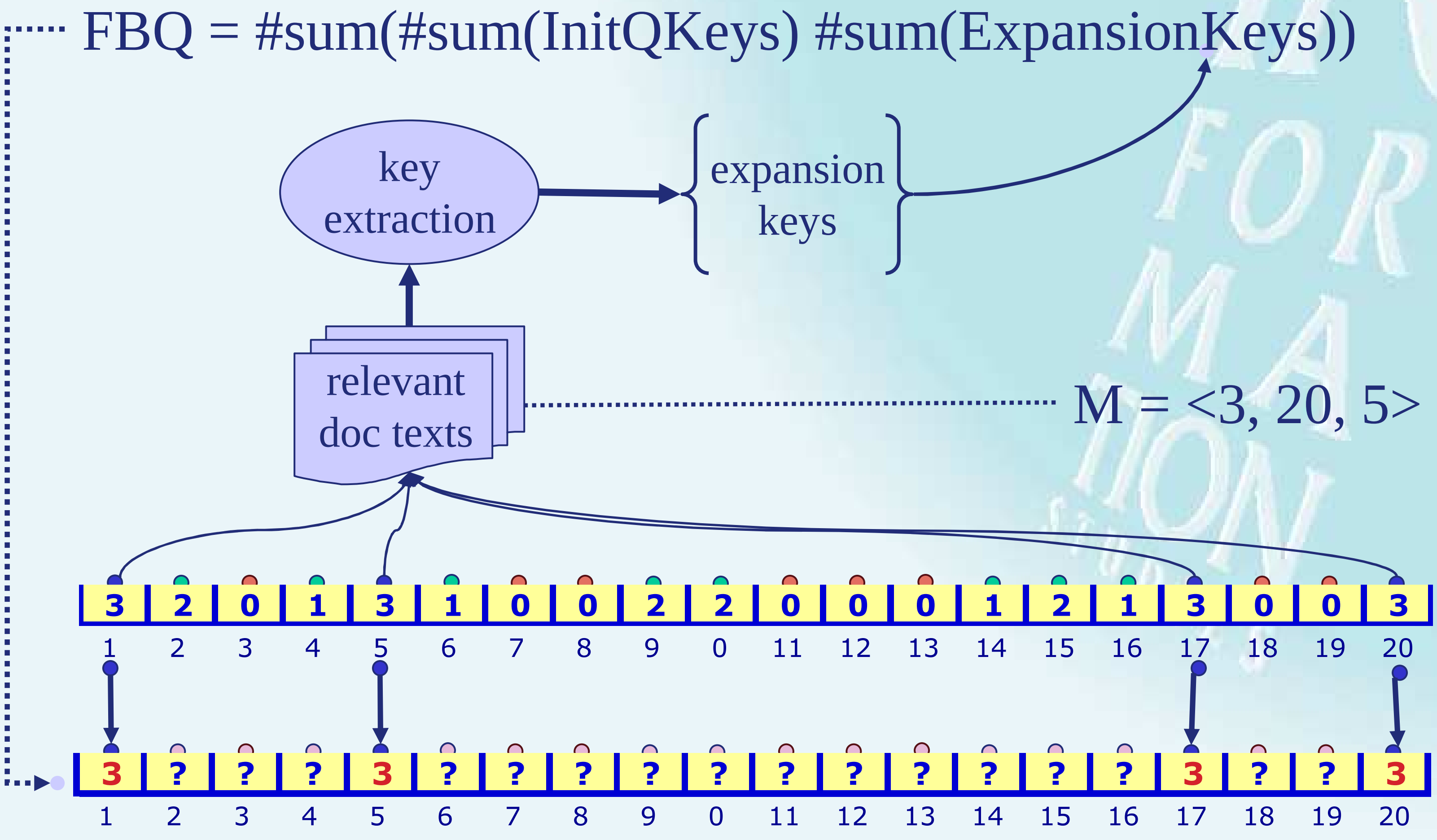
Browse window **B**	Feed-back set **F**	**Stringent feedback criterion R=3**	Diff. to baseline (% units)	**Regular feedback criterion R≥2**	Diff to baseline (% units)	**Liberal feedback criterion R≥1**	Diff. to baseline (% units)
30	30	**26.5**	+5.8	**29.5**	+8.8	**30.2**	+9.5
30	10	**26.5**	+5.8	**29.4**	+8.7	**30.1**	+9.4
30	5	**26.6**	+5.9	**28.6**	+7.9	**28.7**	+8.0
30	1	**24.4**	+3.7	**24.2**	+3.5	**24.0**	+3.3
1	1	**21.6**	+0.9	**22.5**	+1.8	**22.9**	+2.2

Stringent Evaluation (MAP (%), N=41, TREC). Baseline MAP=20.2%

Liberal RFB spoils the effect

Browse window B	Feed-back set F	**Stringent feedback criterion R=3**	Diff. to baseline (% units)	**Regular feedback criterion R≥2**	Diff to baseline (% units)	**Liberal feedback criterion R≥1**	Diff. to baseline (% units)
30	30	**37.5**	+17.3	**27.1**	+6.9	**24.9**	+4.7
30	10	**37.5**	+17.3	**27.1**	+6.9	**24.9**	+4.7
30	5	**36.9**	+16.7	**27.5**	+7.3	**23.9**	+3.7
30	1	**31.7**	+11.5	**23.3**	+3.1	**22.6**	+2.4
1	1	**20.8**	+0.6	**21.6**	+1.4	**22.0**	+1.8

Feedback with Freezing Ex.



Ranked RB query result with seen relevant documents frozen to their ranks

Stringent Evaluation (MAP (%), N=41, TREC). Baseline MAP=20.2%

Browse window **B**	Feed-back set **F**	**Stringent feedback criterion R=3**	Diff. to baseline (% units)	**Regular feedback criterion R≥2**	Diff to baseline (% units)	**Liberal feedback criterion R≥1**	Diff. to baseline (% units)
30	30						
30	10						
30	1	**27.7**	+7.5				
30	1	**31.7**	+11.5				
1	1						

with freezing

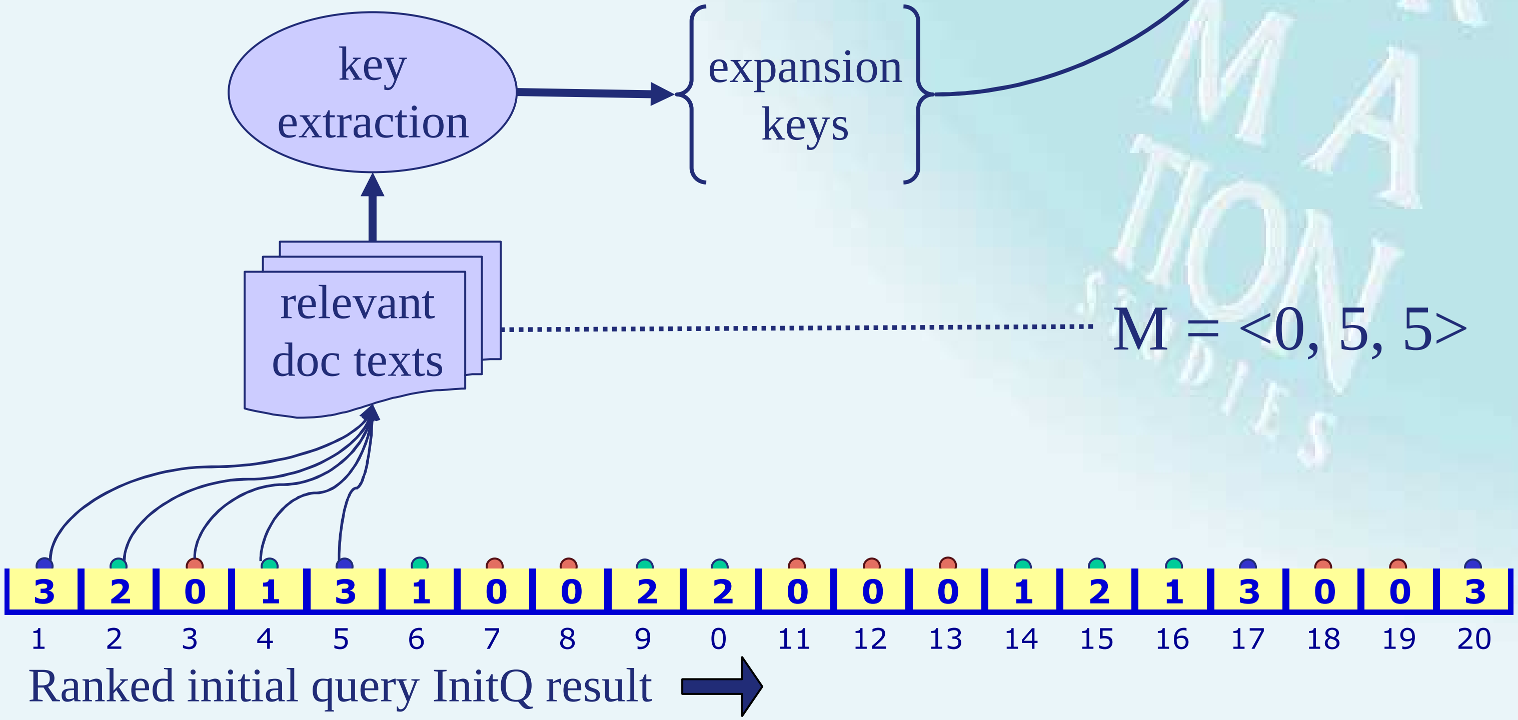
no freezing

Pseudo Relevance Feedback

- No user interaction after initial search
- The first *N* results are assumed relevant
- Their index features are used to revise the original query
- Evaluated PRF at three relevance thresholds (stringent, regular, liberal)

Pseudo RFB Model Example

FBQ = #sum(#sum(InitQKeys) #sum(ExpansionKeys))



Pseudo RFB

- 1, 5, 10 or 30 top-docs used for feedback; baseline MAP 20.2%
- Stringent, regular and liberal evaluation levels
- Results clearly dependent on evaluation stringency

PRF Top B docs	**PRF MAP (%)** Stringent Eval	Diff. to baseline (% units)	**PRF MAP (%)** Regular Eval	Diff to baseline (% units)	**PRF MAP (%)** Liberal Eval	Diff. to baseline (% units)
30	**19.8**	-0.4	**25.1**	+2.4	**24.2**	+3.5
10	**19.5**	-0.7	**25.8**	+3.1	**24.5**	+3.8
5	**21.2**	+1.0	**25.8**	+3.1	**24.1**	+3.4
1	**22.0**	+1.8	**25.3**	+2.6	**22.8**	+2.1

Conclusions

When liberal relevance threshold is used in evaluation:

- A small number of highly relevant FB documents <u>did not</u> outperform several mixed quality FB documents
- Even if high-quality FB would be given, its effects remain unseen

Conclusions 2

When stringent evaluation criterion is used in evaluation:

- relevance threshold for RF documents should be kept high
- lots of mixed quality RF documents distort the RFB effect of highly relevant documents among them

Conclusions 3

- PRF improved effectiveness when liberal evaluation criteria were used - by a typical percentage - but not with stringent evaluation
- PRF adds marginal documents – is this what we want?
- Are we missing possibilities for useful query/system adaptation?

Further Work

- 1. How can negative RFB be applied in interactive IR with graded judgments?
- 2. How effective is negative RFB in a graded assessment environment?
- 3. How does user's domain knowledge / consistency affect RFB
- An extended user model developed

Thank you!