search engine group

RMIT UNIVERSITY Computer Science

When Query Expansion Fails

Bodo Billerbeck

Justin Zobel

School of Computer Science and Information Technology, RMIT University

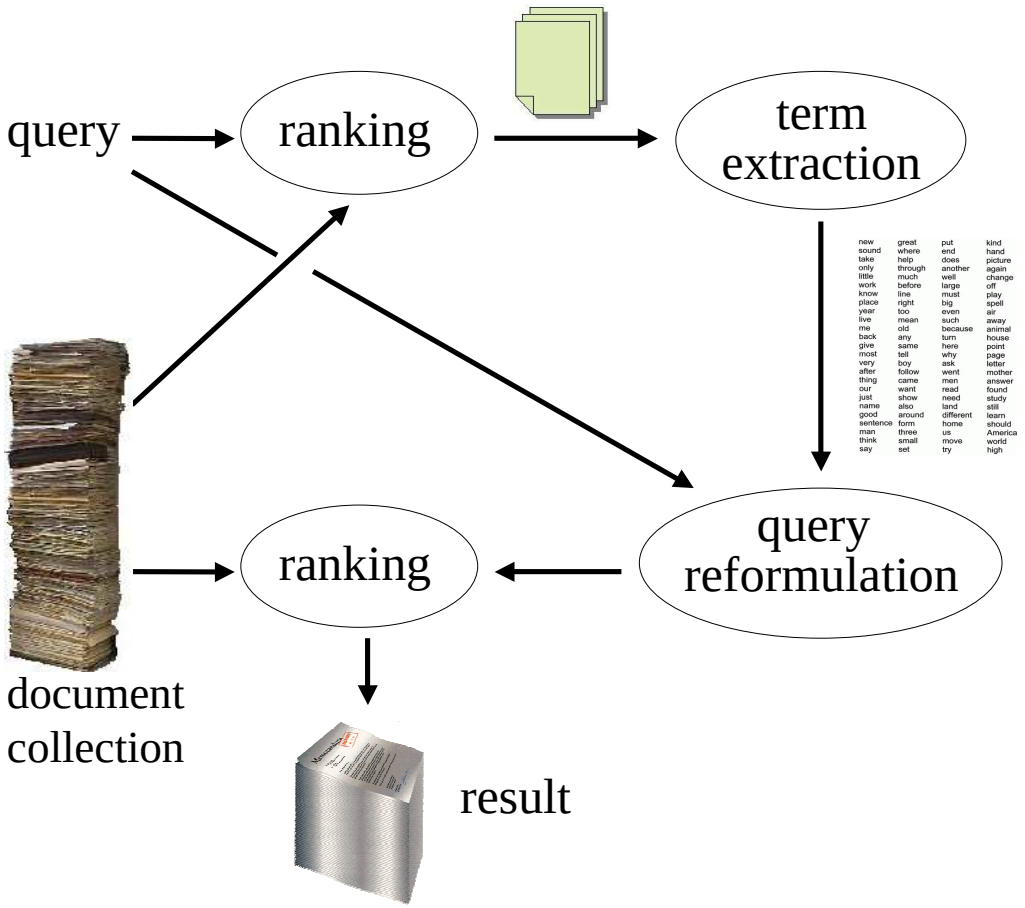
Query expansion

"Blind Relevance Feedback" (as described by Robertson & Walker)

- **Step 1**: Documents are ranked using the original query
- **Step 2**: Using the term selection formula, the best E terms are selected from the top R documents

$$TermSelectionValue = \left(\frac{n_t}{N}\right)^{r_t}\binom{R}{r_t}$$

  - n_t: number of documents containing term t
  - N: number of documents (e.g., 523,587 for TREC 8)
  - R: number of (assumed) relevant documents
  - r_t: number of relevant documents containing term t
- **Step 3**: Expansion terms are assigned relevance values and added to the original query
- **Step 4**: Documents are ranked using the reformulated query



Successful & unsuccessful expansion

search engine group

RMIT UNIVERSITY Computer Science

TREC query 405: "cosmic events"

Expansion terms:

asteroid, asteroids, astronomers, astronomical, astronomy, cobe, comet, comets, cosmic, cosmological, data, dust, earth, events, issledovaniya, jovian, kosmicheskiye, nasa, orbits, particles, planet, scientists, solar, space, spaceguard, sunless, telescope

	raw	expanded
Recall (out of 38)	13	31
Av. Precision	0.0612	0.2158
P@ 5 docs:	0.4000	0.4000
P@ 10 docs:	0.3000	0.5000
P@ 15 docs:	0.2667	0.3333
P@ 20 docs:	0.2000	0.4000
P@ 30 docs:	0.1667	0.3667
P@ 100 docs:	0.0700	0.2100
P@ 200 docs:	0.0500	0.1200
P@ 500 docs:	0.0240	0.0560
P@ 1000 docs:	0.0130	0.0310

TREC query 440: "child labor"

Expansion terms:

age, ballboys, batboys, child, clac, detrimental, dol, employed, employers, employment, flsa, hazardous, hours, labor, minors, nonagricultural, nonschool, occupational, occupations, olds, permissible, reg, school, subpart, wecep, workers, young

	raw	expanded
Recall (out of 54)	30	5
Av. Precision	0.0925	0.0033
P@ 5 docs:	0.2000	0.0000
P@ 10 docs:	0.2000	0.0000
P@ 15 docs:	0.2000	0.0667
P@ 20 docs:	0.2500	0.0500
P@ 30 docs:	0.2000	0.0333
P@ 100 docs:	0.1500	0.0200
P@ 200 docs:	0.1050	0.0150
P@ 500 docs:	0.0560	0.0100
P@ 1000 docs:	0.0300	0.0050

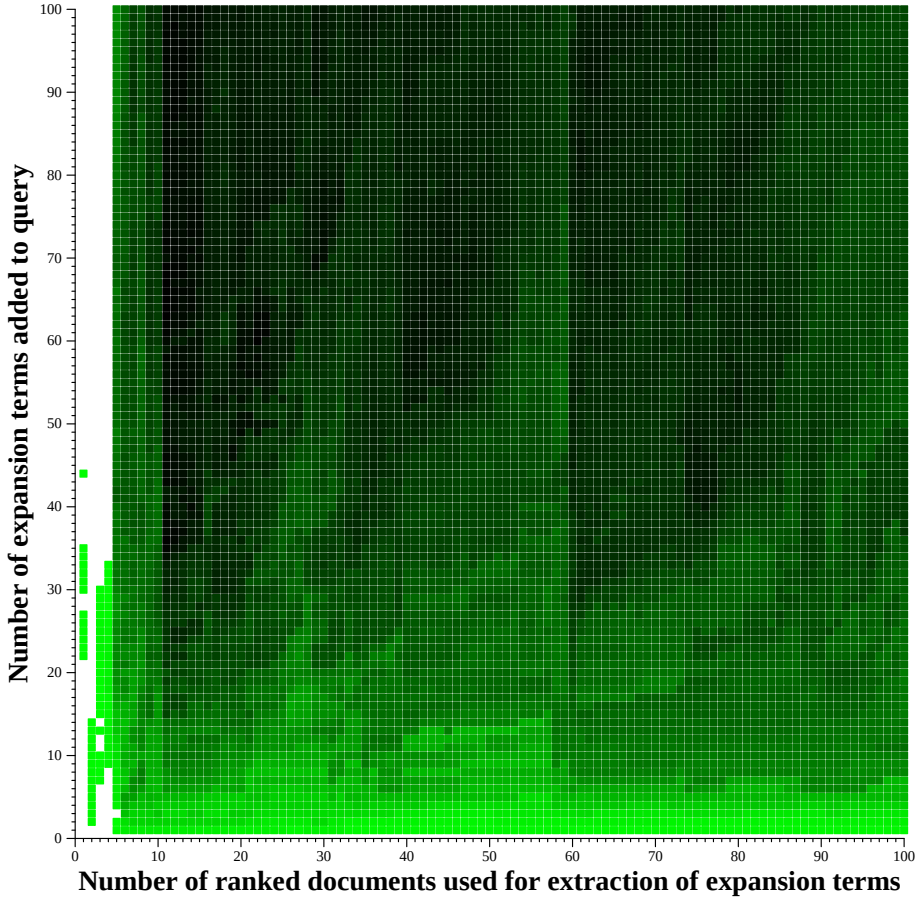
Is this QE robust? Examining parameters

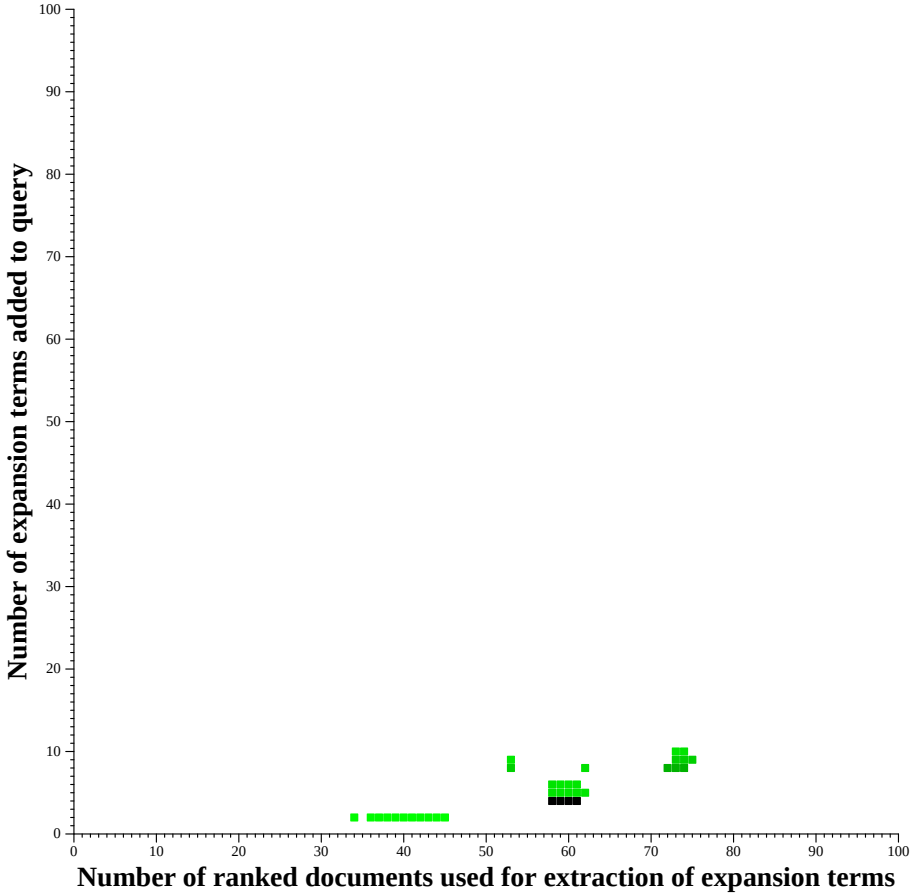
search engine group

RMIT UNIVERSITY Computer Science

- Two main parameters are examined:
  - number of documents (R) from which expansion terms are sourced
  - number of expansion terms (E) appended to a query
- There is no well found basis for choosing parameters for a particular query/collection. Although one method uses thresholding to choose E, we found that this does not work reliably and, in particular, fails on the web TREC data.
- When examining what happens when changing those two parameters, we found that the average precision of nearly all queries changed. Some got better, others got worse. In some cases there were large jumps due to just one document added to the list from which expansion terms were sourced.

TREC Query 405 versus 440

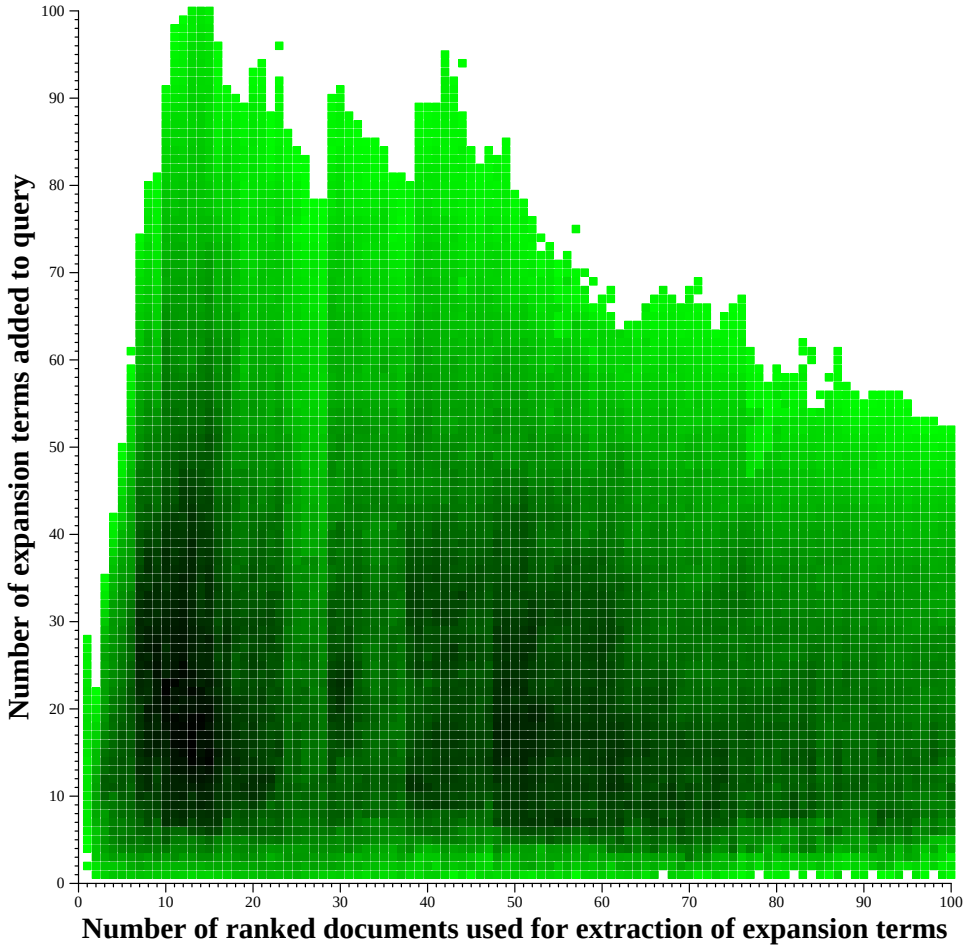




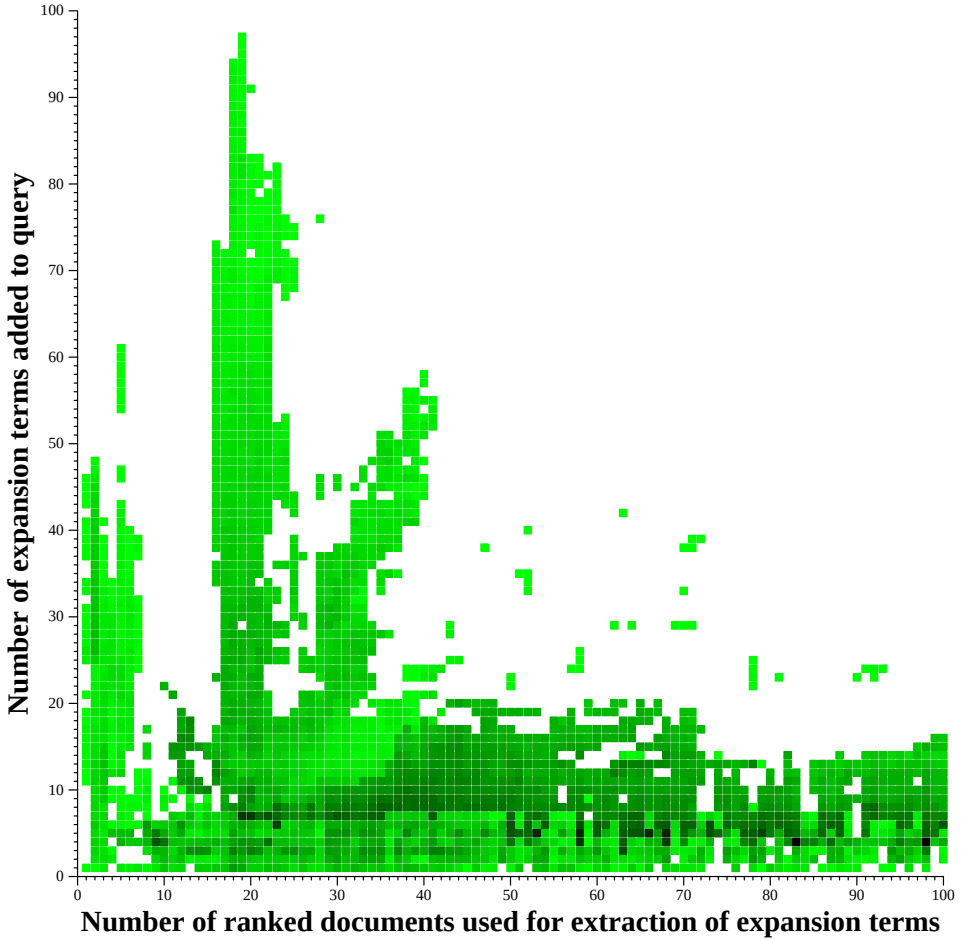
- The darker the dot, the higher is the average precision for that parameter pair
- A vast range of parameters lead to good improvement of average precision

- No colour means that average precision for that combination is worse than no expansion
- QE fails for all but a very small number of R and E pairs

Best R, t for 'average' query

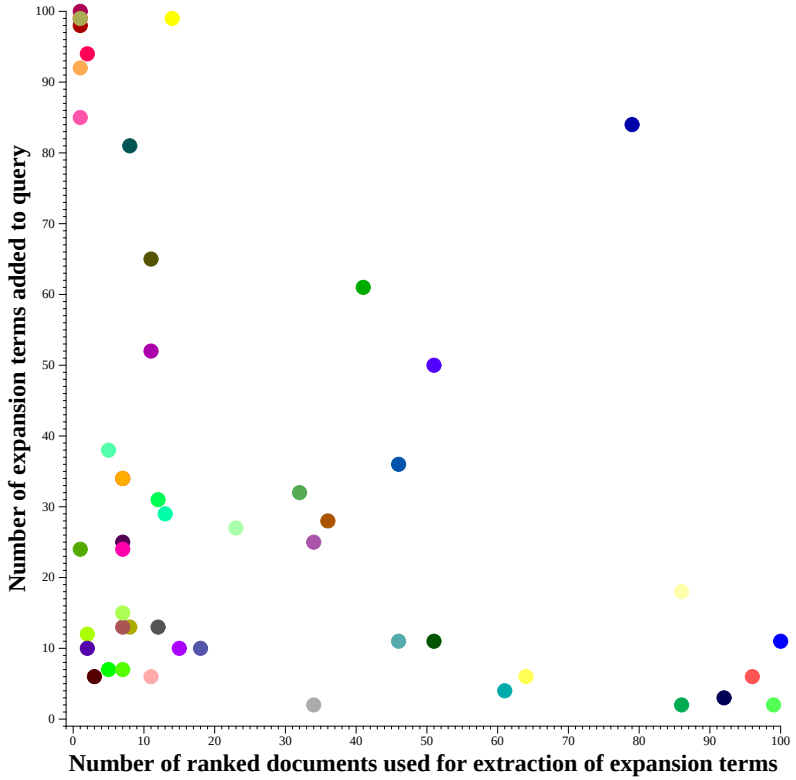


- Many parameter combinations improve average precision for TREC 8 (newswire)
- Best combination: R = 13; E = 15;

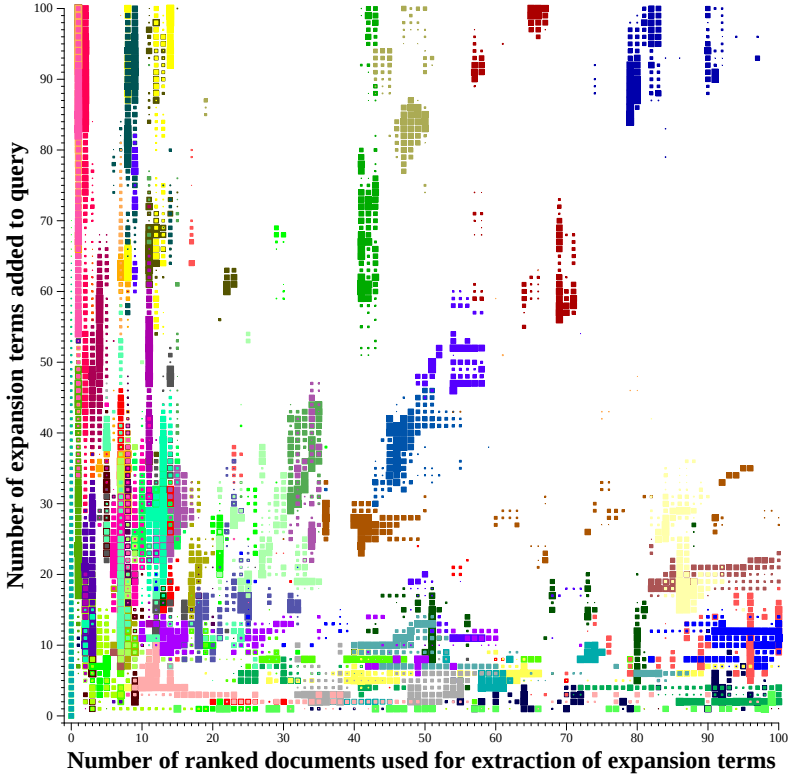


- Few combinations lead to an increase in average precision for TREC 9 (web)
- Best combination: R = 98; E = 4;

Best R, E for each of 50 queries

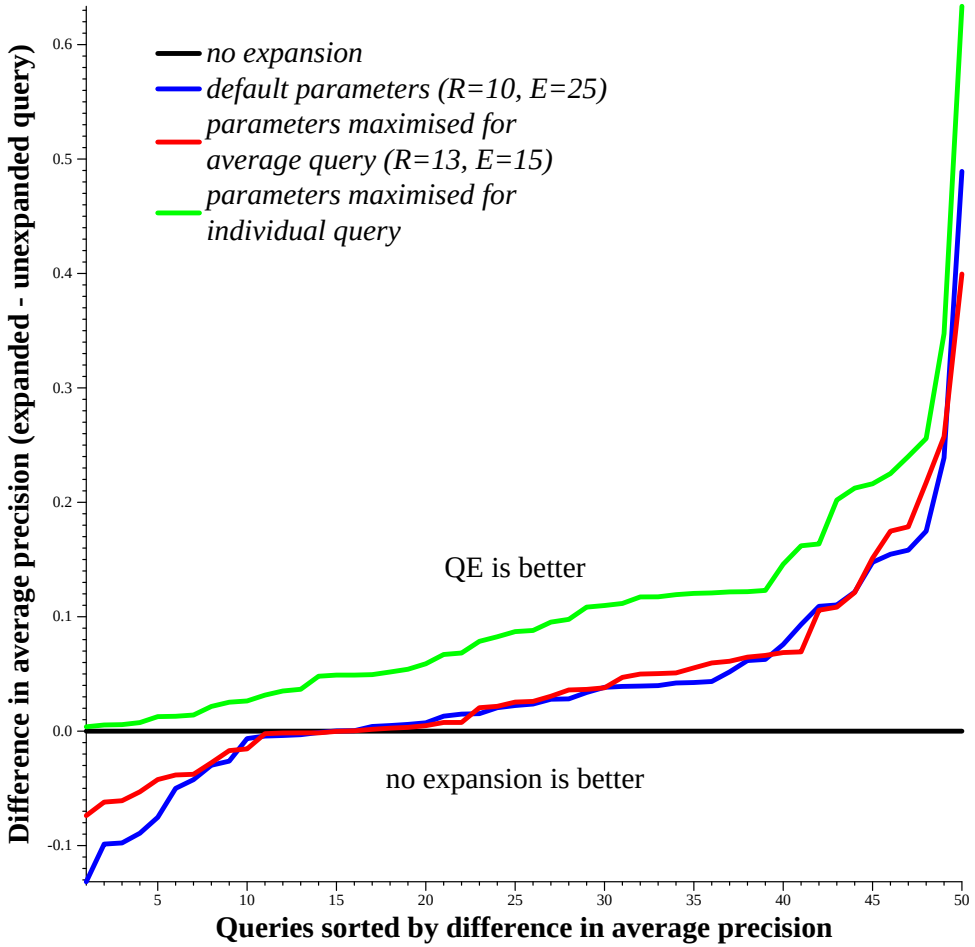


- The best parameter combination is distinctly different for each query.
- Some queries are best expanded by adding a large number of terms from only few docs, others work better when adding only few terms from a large number of docs.

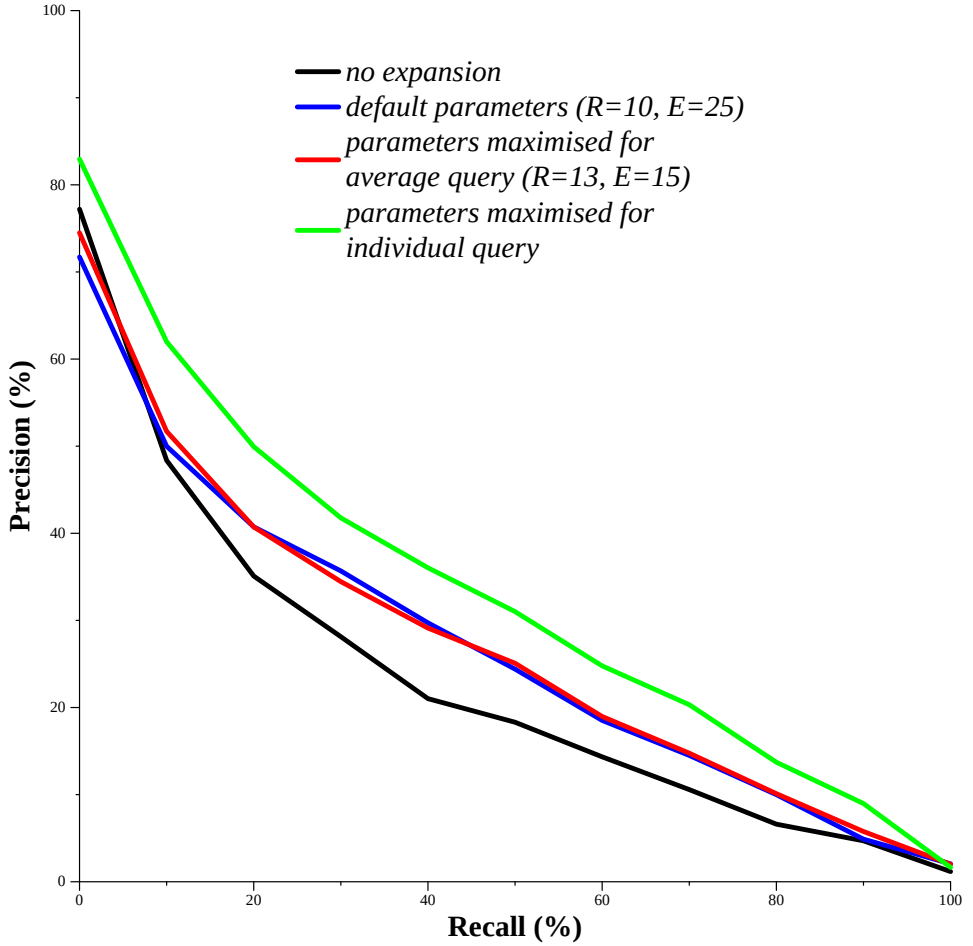


- Best 100 combinations per query.
- For some queries a particular document contains many relevant terms (vertical streaks).
- For others a particular term leads to the best improvement (horizontal streaks).

Average precision with best parameter pair



- Number of queries that are degraded/had no change/are improved through expansion.



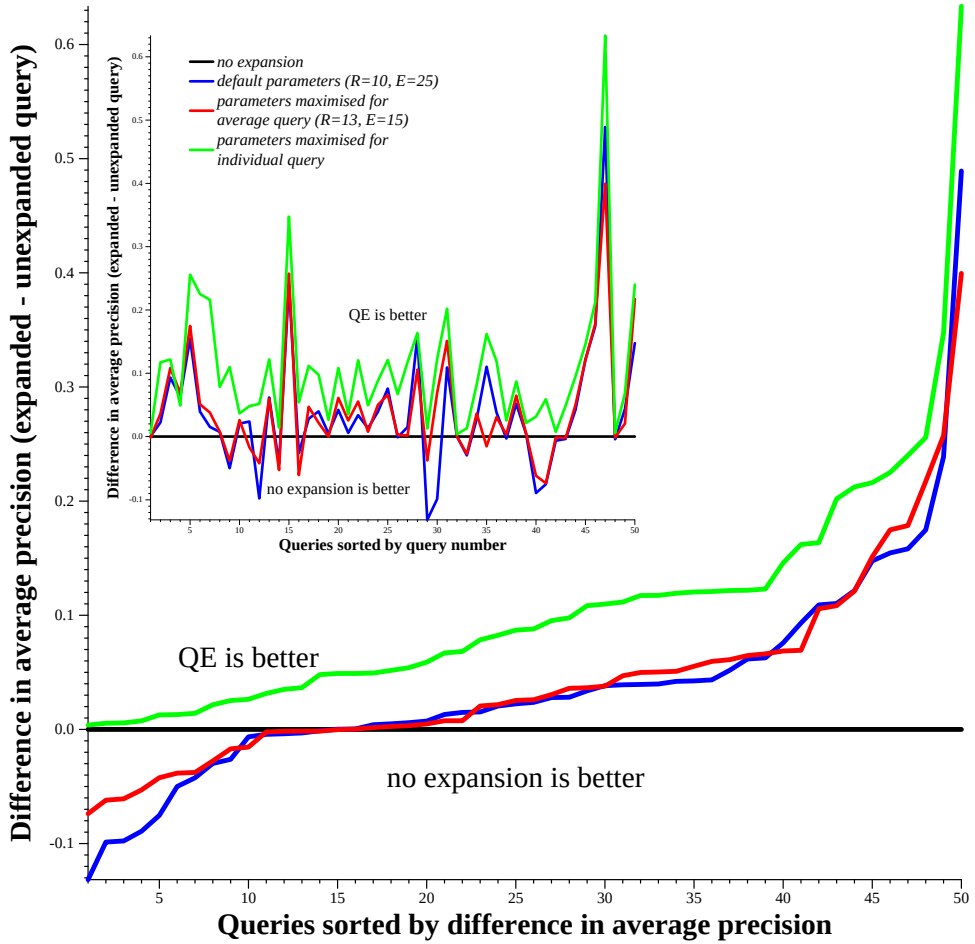
- The recall/precision graph shows by how much maximising average precision improves effectiveness

What have we tried? What can we learn?

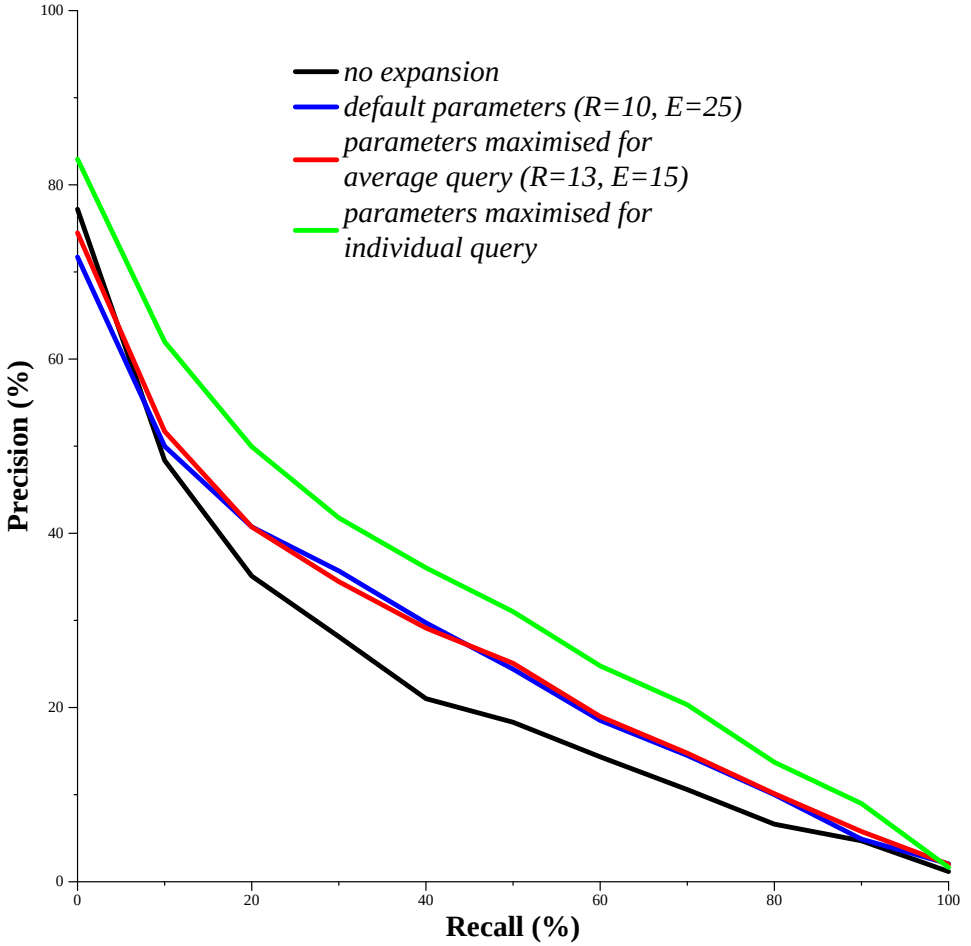
- We have investigated potential correlations between different statistics, but none were significant (using the signed rank test). Correlations included:
  - Average precision increase versus average document score of documents used for expansion, where the score is calculated as follows (using 0.9 million unique excite queries): +10 for first rank, ... +1 for 10th rank, and +0 otherwise
  - Average precision increase versus average accumulator score of documents in R (testing the default (R=10, E=25), best average (R=13, E=15), individual maximum)
  - Increase in average precision versus okapi weight of indexed, non-stopped query terms
  - We also tried various other document scores, such as the sum over all term weights of the terms occurring in the set of R documents, where the weight is calculated by using their tf-idf values.

- For which pair of parameters does QE work?
  - There is no single parameter pair that improves average precision for all queries, let alone maximises average precision for all queries.
  - There is no single parameter pair that improves average precision averaged over a large number of queries for all collections (or even any two collections, as it seems).

Average precision with best parameter pair



- Number of queries that are degraded/had no change/are improved through expansion.



- The recall/precision graph shows by how much maximising average precision improves effectiveness