Query characterization and performance prediction

Design notes for ☑ ~~NDS-919~~ - Query characterization RESOLVED and ☑ ~~NDS-920~~ - **Investigate query performance predictors** RESOLVED .

Query characterization

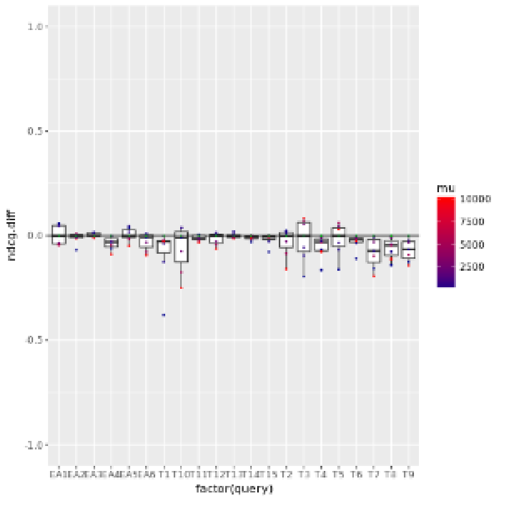
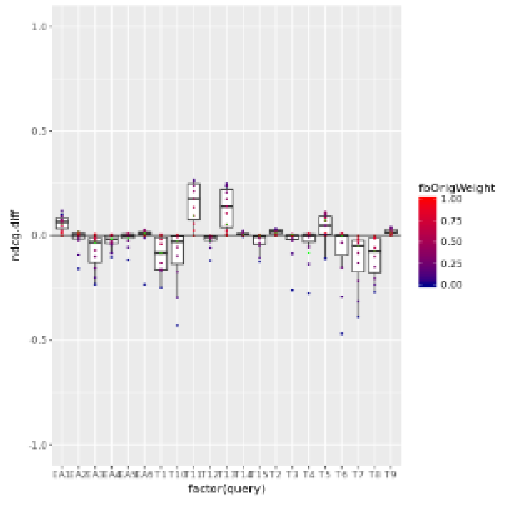
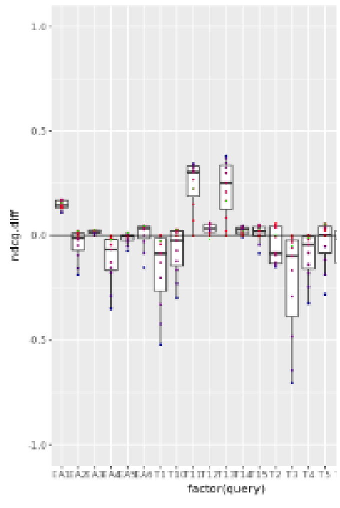
From the BioCADDIE Results, we can see that the PubMed and Wikipedia expansion models provide some improvement, but not at the higher ranks. As is often the case with expansion, inspection of individual queries shows that while some queries benefit from expansion others do not.

The following plots illustrate the effect of varying the Dirichlet mu parameter (a), RM3 fbOrigWeight parameter for RM3 (b), PubMed RM3 (c), and Wikipedia RM3 (d) for NDCG@1000 and NDCG@20. The zero line is the per-query cross-validated QL (Dirichlet) score for each topic. The box plots represent the variation in scores as the parameter changes. For the RM3 models, all other parameters are fixed at their cross-validated values (i.e., only fbOrigWeight is changed). The X axis shows the BioCADDIE topics, the Y-axis is the difference in NDCG from the cross-validated QL/Dirichlet baseline. The boxplots represent one point per parameter value – blue for low and red for higher parameter values. The green dots are the cross-validated fbOrigWeight values for the RM models.

NDCG@1000

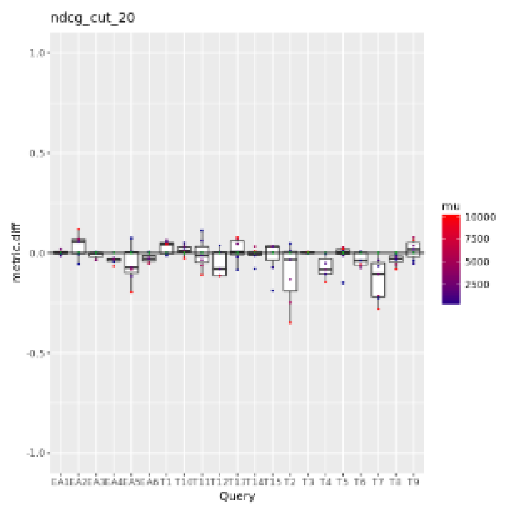
Looking at a few examples:

- In plot (a) we see that NDCG decreases for topic T10 as mu increases, but for T13 varying mu has little effect.
- In plot (b) we see that decreasing the fbOrigWeight (and therefore increasing the effect of the expansion terms) has a positive effect for topics T11 and T13, but a negative effect for topic T7.
- In plot (c) we see that PubMed expansion has a negative effect for T3 as compared to RM3 expansion. PubMed expansion is riskier than RM3 expansion.
- Similarly, in plot (d) we see that Wikipedia expansion in general has a negative effect compared to RM3 or PubMed RM3 expansion.

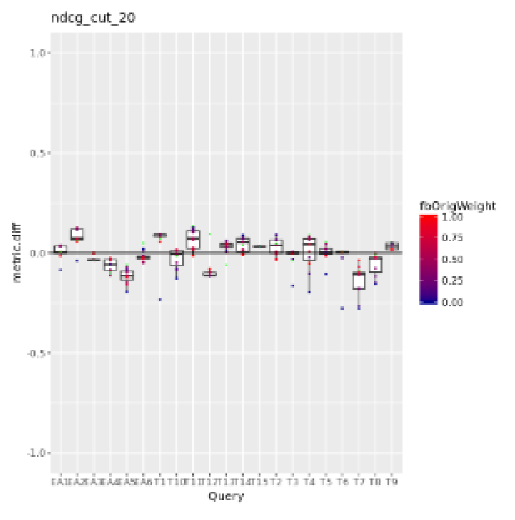
		
(a) Dirichlet (mu)	(b) RM3 (fbOrigWeight)	(c) PubMed RM3 (fbOrigW

NDCG@20

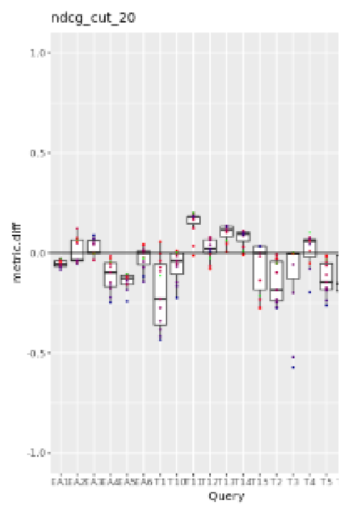
Looking instead a the higher ranks, we see a much more muted effect of expansion with the BioCADDIE test collection. As above, expansion appears to be effective for a small number of test queries (< 0.5).



(a) Dirichlet (mu)



(b) RM3 (fbOrigWeight)



(c) PubMed RM3 (fbOrigWeight)

In our current runs, for each held-out query the fbOrigWeight that controls the mixing of the original and feedback query is learned from the training queries. The values are relatively fixed. In the next section, we explore whether we can reliably predict when to apply one model or another or to predict the fbOrigWeight mixing parameter via query performance prediction methods.

Query performance prediction

A central goal of query performance prediction or query difficulty estimation is to identify features, both pre- and post-retrieval, that can be used to predict the performance of a query. A common approach is to predict some effectiveness metric (e.g., MAP/NDCG). In the past, performance predictors were evaluated based on simple correlation with the target metric. Hauff et al (2009) argue that linear correlation coefficients are misleading and overstate performance. They instead focus on RMSE (through linear regression) for comparison of individual predictors and penalized regression for combinations of predictors.

For BioCADDIE, we are focused on expansion models and therefore are primarily concerned with adaptive feedback. Lv and Zhai's (2009) approach seems to be the most applicable – estimating the feedback mixing parameter per-query. This will require the following:

- A framework for implementing baseline and custom predictors (ir-utils or otherwise)
    - Preliminary implementation in edu/gslis/biocaddie/qpp/predictors
- Ability to generate a set of pre- and post-retrieval predictor values for each query for multiple collections. This will output a matrix of queries and predictors.
    - edu.gslis.biocaddie.qpp.RunQPP produces predictor matrix.
- Calculate correlation (Pearson and Spearman), RMSE between the predictor and a given metric or parameter (eg., RM3 lambda)
- Ability to select features (manually or automatically) and to construct a predictive model (i.e, regression) using one or more predictors.
    - Investigating penalized regression via R glmnet.
- Evaluate the predictive model via cross validation.
    - Implemented preliminary draft of CrossValidateQPP class.

Adaptive feedback

The approach explored by Lv and Zhai (2009) is to learn a model to predict the expansion mixing weight. They found six features to be predictive of the feedback weight in a linear model (in order of significance). All of these predictors are post-retrieval predictors, some with significant overhead (marked with *).

- Topic model clarity (FBEnt_R3*): Relative entropy of the feedback document topic model to the collection.
- Exponentiated feedback clarity (FBEnt_R2*): Exponentiated relative entropy of feedback documents to the collection
- Divergence (QFBDiv_A): KL-divergence of query and feedback documents
- Feedback radius (FBRadius): average divergence between each document and the centroid of the feedback documents.
- Query clarity (QEnt_R1): Relative entropy of the query compared to the collection
- Log query clarity (QEnt_R3): Log of the relative entropy of the query compared to the collection

There are a variety of other predictors, such as those discussed in Carmel and Yom-Tov's (2010) monograph. These include:

Pre-retrieval predictors:

- IDF (mean, min, max, variance)
- Inverse collection term frequency (ICTF: mean, min, max, variance)
- Collection query similarity (SCQ, Zhao et al)
- Simplified clarity score (SCS)
- Predictors requiring additional computation
    - Coherency (He et al)

    - PMI/avgPMI/maxPMI
    - Term-weight variability (maxVar) - Zhao et al, also Hauff.

Post-retrieval predictors:

- Clarity (Cronen-Townshend, 2002)
- Drift (Shtok, Kurland, Carmel, 2009)
- Deviation (Perez-Iglesia and Araujo, 2010)

Preliminary results

References

Carmel, D., & Yom-Tov, E. (2010). Estimating the Query Difficulty for Information Retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *2*(1), 1–89. http://doi.org/10.2200/S00235ED1V01Y201004ICR015

Cronen-Townsend, S., Zhou, Y., & Croft, W. B. (2002). Predicting Query Performance. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 299–306). New York, NY, USA: ACM. http://doi.org/10.1145/564376.564429

Hauff, C., Azzopardi, L., & Hiemstra, D. (2009). The Combination and Evaluation of Query Performance Prediction Methods. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval* (pp. 301–312). Berlin, Heidelberg: Springer-Verlag. http://doi.org/10.1007/978-3-642-00958-7_28

Lv, Y., & Zhai, C. (2009). Adaptive Relevance Feedback in Information Retrieval. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 255–264). New York, NY, USA: ACM. http://doi.org/10.1145/1645953.1645988

Zhao, Y., Scholer, F., & Tsegay, Y. (2008). Effective Pre-retrieval Query Performance Prediction Using Similarity and Variability Evidence. In C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, & R. W. White (Eds.), *Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings* (pp. 52–64). Berlin, Heidelberg: Springer Berlin Heidelberg. http://doi.org/10.1007/978-3-540-78646-7_8