Learning to Rank System Configurations

ABSTRACT

Information Retrieval (IR) systems heavily rely on a large number of parameters, such as the retrieval model or various query expansion parameters, whose values greatly influence the overall retrieval effectiveness. However, setting all these parameters individually can often be a tedious task, since they can all affect one another, while also vary for different queries. We propose to tackle this problem by dealing with entire *system configurations* (i.e. a set of parameters representing an IR system) instead of single parameters, and to apply state-of-the-art Learning to Rank techniques to select the most appropriate configuration for a given query. The experiments we conducted on two TREC AdHoc collections show that this approach is feasible and significantly outperforms the traditional way to configure a system, as well as the top performing systems of the TREC tracks. We also show an analysis on the impact of different features on the model's learning capability.

Categories and Subject Descriptors

H.3.3 [**Information Storage & Retrieval**]: Information Search & Retrieval

Keywords

Information retrieval, learning to rank, retrieval system parameters

1. INTRODUCTION

The effectiveness of Information Retrieval (IR) systems heavily relies on a large number of parameters, ranging from the choice of stemmer, the smoothing technique to the number of terms that are added to the query when performing automatic query expansion. Over the years, and through the evaluation forums such as TREC, the IR community has produced abundant field knowledge, scattered in the literature, on setting the appropriate values of these parameters, in order to optimise the performance of the retrieval systems. The parameters are usually studied in isolation: one attempts to set the value for only one or a few parameters at a time, without taking into account the influence of the setting of other parameters. This makes it difficult to configure a globally optimised set of parameters, as modern IR systems involve a quite large number of parameters that are mutually dependent. There has been no systemic study trying to set the best values for all the parameters of a system.

In this paper, we treat a complete set of parameters of a system as a *system configuration*. We cast the problem of selecting the most appropriate system configuration as a configuration ranking problem using a learning to rank (L2R) [7] approach. The candidate space is formed of tens of thousands of possible system configurations, each of which sets a specific value for each of the system parameters. L2R models are trained to rank them with respect to a performance measure. This approach has the advantage of taking all the system parameters into account at the same time, thus allowing them to influence each other. Moreover, our approach can make a query-dependent choice of system configuration, i.e. different search strategies could be selected for different types of query.

The main contribution of this paper is that we propose a new way based on L2R to set system parameters and show its feasibility and competitive or superior retrieval effectiveness to the state of the art on two TREC collections.

2. LEARNING TO RANK SYSTEM CONFIGURATIONS

Our method is based on L2R approaches for IR [7]. However, instead of ranking documents for a query, we rank system configurations. The problem is formulated below.

We assume that an IR system involves a set $\mathbf{P}$ of parameters. Each parameter $p_i \in \mathbf{P}$ can take a value from its domain D_i. Therefore, we have $\prod_i |D_i|$ possible configurations (without considering the fact that some configurations are impossible). This number could be very large, given the quite large number of system parameters used in modern systems and their possible values. We also assume that we have a set $\mathbf{Q}$ of queries for which we have relevance judgments on a document collection, which can be used to generate training examples for L2R models: for each possible configuration $c_j \in \mathbf{C}$, where $\mathbf{C}$ is the space of all configurations, we generate a measure in the IR performance metric (such as MAP, precision, etc.) for the pair (q_k, c_j). Our goal is to rank the possible configurations for a new query q such that the best ranked system configuration could lead to the best performance.

A L2R model is based on a set of features defined on (q_k, c_j). Usually, in L2R approaches, the features are related

to the query, the document (and sometimes the document collection), as well as the relationships between them. In our case, we define features relating to the query q_k and to the configuration c_j. It can be costly to define features on the relationships between the query and the configuration because this requires to run a retrieval operation for each system configuration and to extract some features according to the retrieved results. In a realistic setting, this would be prohibitively expensive. We leave the problem of defining such features to our future work. Our primary goal in this paper is to test the feasibility of L2R approaches for setting system configurations.

In this work, we consider a set of common system parameters (see Table 1): 1 parameter for retrieval model and 4 parameters for pseudo-relevance feedback. Notice that these parameters are all for retrieval (not for indexing) and can be set for a query on the fly. For each of the possible configurations [1], we run the Terrier IR system [10] to obtain the corresponding performance measures. This amounts to more than 10,000 different configurations for each training query.

Table 1: Description of the system parameters that we use to build our dataset

Parameter	Description & values [2]
Retrieval model	21 different retrieval models: DirichletLM, JsKLs, BB2, PL2, DFRee, DFI0, XSqrAM, DLH13, HiemstraLM, InL2, DLH, DPH, IFB2, TFIDF, InB2, InexpB2 , DFRBM25 , BM25, LGD, LemurTFIDF, InexpC2.
Expansion model	7 query expansion models: nil, Rocchio, KL, Bo1, Bo2, KLCorrect, Information, KLComplete.
Expansion documents	Number of documents used for query expansion: 2, 5, 10, 20, 50, 100.
Expansion terms	Number of expansion terms: 2, 5, 10, 15, 20.
Expansion min-docs	Minimal number of documents an expansion term should appear in: 2, 5, 10, 20, 50.

We define a set of 65 features for L2R models on a pair (q_k, c_j) that can be divided into four different groups: 43 features computed using query word statistics (QueryStats), 16 features describing the linguistic properties of the query (QueryLing), 1 feature describing the retrieval model (RetModel), and 4 features related to query expansion parameters (Expansion). The two last groups of features are the same as shown in Table 1. We provide a brief description of QueryStats and QueryLing features.

QueryStats: These features are query-dependent statistical features that were previously used in both query difficulty prediction [3, 5] and learning to rank [8] settings. These features include query terms statistics such as variations of their IDF in the collection (min, max, avg, and sum IDF over the query terms), Query Feedback [13] (min, max, etc.) calculated using various numbers of feedback documents and several default retrieval models such as QL and Bo2, or variant of the NQC which is based on the standard deviation of retrieved documents scores [11].

QueryLing: These are also query-dependent features, but they focus on modelling the linguistic properties of the query. We implemented the features defined in [9], such as the number of WordNet synsets for query terms, the number of prepositions in the query, and so on.

Query-dependent features aim to inform the L2R technique about the characteristics of the query, thus allowing to select different system configuration on a per query basis. As a first investigation, we use all the reasonable features at our disposal without performing any feature selection, leaving this aspect for future work.

We use three common performance metrics to rank system configurations: MAP, P@100, and Rprec. These metrics are chosen because they were found to be the least correlated [1].

Three types of L2R techniques have been proposed in the literature based on point-wise, pair-wise and list-wise principles [7]. The point-wise approaches aim at learning to predict a relevance score or class for each document, while the pair-wise approaches learn to predict if one document is more relevant than another. Finally, the list-wise models consider the whole list of documents and optimise a ranking measure. All the L2R models could be suitable to our task: they can rank system configurations in such a way that the best configuration will be ranked first. This is the configuration that we want to select. Notice, however, that the relative positions of the elements at lower positions are also important in L2R models, in particular in pair-wise and list-wise models. The optimisation related to this part of ranking may not be crucial or necessary for our task. Our learning objective could be different. However, we do not examine this question in this paper.

We experimented with a large selection of the existing L2R techniques made available by the RankLib[3] and the SVM^{rank} [4] toolkits. The performance varies largely among the models. Due to space limit, we only report the results with the following representative models: Gradient Boosted Regression Trees (GBRT) [4], Random Forests [2], LambdaMART [12], and SVM^{rank} [6], since we found that these techniques were the most effective for our task. Our selection of L2R techniques covers all the three categories: GBRT and Random Forests are point-wise techniques, SVM^{rank} is pair-wise, and LambdaMART is classified as both pair-wise and list-wise. We trained the three techniques implemented in RankLib to optimise nDCG@10, while SVM^{rank} has its own optimisation criteria.

3. EXPERIMENTAL RESULTS

We carried out experiments on two TREC AdHoc test collections: TREC-7 (queries 351-400) and TREC-8 (queries 401-450) for which we merged the queries into one set since the document collection is the same.

We used a 5-fold cross validation, where each fold has separate training (3/5), validation (1/5), and test sets (1/5). The training queries are used to train L2R models, the validation queries are used to minimise over-fitting, and the test queries are used to evaluated the learned models. We report the average performance on the test queries in Table 7. For each test query, we use the system configuration that has

[1]Impossible configurations, such as using nil expansion method but a number of expansion documents and terms, are excluded.

[2]Details can be found at `terrier.org/docs/v4.0/javadoc`.

[3]`sourceforge.net/p/lemur/wiki/RankLib/`

[4]`www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

Table 2: Results with different L2R models and feature abalations. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates figure Statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP		P@100		RPrec	
BM25	0.1942		0.1719		0.2330	
Grid Search	0.2480		0.2213		0.2835	
Random Forests (All)	0.3319 △		0.2785 △		0.3439 △	
- QueryStats	0.3180 △	(-4.17%)	0.2947 △	(+5.80%)	0.3658 △	(+6.35%)
- QueryLing	0.3367 △	(+1.43%)	0.2835 △	(+1.80%)	0.3507 △	(+1.96%)
- RetModel	0.3210 △	(-3.28%)	0.2746	(-1.44%)	0.3462 △	(+0.65%)
- Expansion	0.2201▼	(-33.68%)	0.1843▼	(-33.84%)	0.2384▼	(-30.69%)
SVM^{rank} (All)	0.3073 △		0.2529		0.3204	
- QueryStats	0.2820 △	(-8.23%)	0.2667 △	(+5.48%)	0.3304 △	(+3.12%)
- QueryLing	0.2918 △	(-5.03%)	0.2501	(-1.11%)	0.3498 △	(+9.19%)
- RetModel	0.3118 △	(+1.48%)	0.2628 △	(+3.91%)	0.3400 △	(+6.10%)
- Expansion	0.1723▼	(-43.92%)	0.1203▼	(-52.43%)	0.1914▼	(-40.28%)
GBRT (All)	0.3338 △		0.2803 △		0.3400 △	
- QueryStats	0.3375 △	(+1.11%)	0.2699	(-3.71%)	0.3275 △	(-3.71%)
- QueryLing	0.2982 △	(-10.68%)	0.2908	(+3.75%)	0.3288 △	(-3.31%)
- RetModel	0.3299 △	(-1.17%)	0.2702	(-3.62%)	0.3581 △	(+5.32%)
- Expansion	0.2345▼	(-29.75%)	0.1775▼	(-36.66%)	0.2505▼	(-26.32%)
LambdaMART (All)	0.3271 △		0.2772 △		0.2873	
- QueryStats	0.3272 △	(+0.03%)	0.2705 △	(-2.42%)	0.2692	(-6.28%)
- QueryLing	0.3324 △	(+1.62%)	0.2695 △	(-2.78%)	0.3486 △	(+21.34%)
- RetModel	0.3144 △	(-3.87%)	0.2713 △	(-2.13%)	0.3528 △	(+22.78%)
- Expansion	0.2188▼	(-33.11%)	0.1456▼	(-47.49%)	0.2078▼	(-27.67%)
Upper bound (oracle performance)	0.4136		0.3434		0.4490	

been ranked first by the learned models.

In order to evaluate the effect of each group of features presented in Section 2 (QueryStats, QueryLing, RetModel, Expansion) for selecting the configuration, we perform the following abalation analysis: we remove one group of features at a time, and perform again the learning and testing without the ablated features in order to see the change in retrieval effectiveness. A large decrease in retrieval effectiveness would indicate that the ablated features are deemed important for the learner.

The rows with (All) mean that the models have been learned using the full set of 65 features, while the other rows exhibit results without the ablated group of features. We compare the results of our approach to two baselines: a Grid Search method, which selects the best configuration on a set of training queries (we used both the training and the validation queries here) and uses it on the test queries. This corresponds to the common practice in IR for setting multiple parameters at once. Notice that this configuration is query-independent. For an easy comparison, we also provide the performance of a standard BM25 run (without query expansion), using the default configuration provided by Terrier. We also report in Table 7 the Upper bound of our method, which uses the best possible system configuration for each query (i.e. the oracle performance).

On analysing Table 7, we can make three main observations. Firstly, and most importantly, we see that all L2R techniques can effectively learn to rank reasonable system configurations. All the L2R models can produce much better results than the traditional way (Grid Search) to set system parameters. This result clearly indicates the benefit of using a L2R model to select an appropriate system configuration for a query, rather than setting a unique configuration globally. Among the L2R models, Random Forest, GBRT and LambdaMART produce equivalent performances, while SVM^{rank} performs slightly lower. The superiority of pair-wise and list-wise models over point-wise models cannot be concluded. This observation differs slightly from the traditional utilisations of L2R models where pair-wise and list-wise models are found to perform better than point-wise models [7]. The difference can be explained by the fact that the relative positions of configurations at lower ranks have important impact in pair-wise and list-wise L2R odels, while this is not important for our task.

The L2R models also compare favorably to the best performing systems of the TREC-7 and TREC-8 AdHoc tracks. The best systems at TREC-7 and TREC-8 achieved 0.3032 and 0.3303 in MAP, while the LambdaMART (All) model can produce 0.3148 and 0.3396 in MAP on the two separate sets of queries.

Secondly, we observe that ablating the Expansion group of features always significantly decreases the performance of the learned models, hinting the huge importance of these features for learning an effective model. A possible explanation is that the best parameters for query expansion vary a lot across queries, and the other features are unable to make differences among them. Therefore, in absence of the expansion features, the L2R models cannot make an informed

choice on system configuration.

The ablating of the other groups of features has less impact. When we remove the RetModel feature, the performances can increase or decrease slightly. This highlights the fact that our approach is able to effectively rank system configuration even without using the retrieval model as a feature. A possible explanation is that the other features are often sufficient to determine the best configuration, or the selections of retrieval model across queries are usually consistent so that the feature about the retrieval model does not provide much additional help. Similar observation also holds on the query-dependent features. However, this does not mean that query-dependent features are useless. There could be more questions about the specific features in this group. In fact, several features are redundant. Several others do not seem to be strongly informative. It would be desirable to perform a feature selection to keep a subset of useful features. This will be part of our future work. Notice also that we have two groups of query features. When we remove one of them in our ablation analysis, it is possible that the other group can still provide similar information about the query.

Finally, the results demonstrate the importance of making query-dependent configuration selections. The low impact of query-dependent features in configuration selection does not mean that the final selection of configuration is not query-dependent. For each query, a L2R model always makes a different selection based on the available features. The benefit of query-dependent selection can be best seen by contrasting the L2R models with Grid search, which sets the best query-independent configuration: We can see tat all the query-dependent configuration selections by L2R models are better than Grid search. However, compared to the Oracle performance, we also see that the selections by L2R models are not always the best for each query. There is much room for improvements on this in the future.

Since the Expansion features are shown to be the most important, we conduct a more detailed feature ablation experiment on some individual features of this group. In this experiment, we compare the performance of the models that have been learned with all features to those of the models that have been learned after removing each of the four Expansion features (see Table 1) individually. We observed that the expansion model appears to be the most influential feature for all L2R techniques, which sounds intuitive. Indeed, if a L2R model is not informed of the expansion model in a configuration, the model is unable to make a reasonable choice because there may be a large variation of reasonable configuration selections with different expansion methods.

4. CONCLUSION

In this paper, we proposed a new approach to set system configuration using learning-to-rank methods. We showed that this is a feasible approach, and it can produce superior performance to the state of the art. Our experiments also showed the importance and benefits to make query-dependent configuration setting. The feature abalation analysis showed various impact of different features. In particular, the features about query showed lower impact than we expected. More investigations are needed to fully understand the reason. In this study, we did not make a selection of the features to be used and simply used all the features proposed in the literature that sound relevant. However, we observe that the relevance of some of the features to our task may be low. The features can also be redundant, providing similar information. It will be useful to perform feature selection in the future. In addition, more useful features could be extracted. In particular, the features reflecting the relationships between a query and a configuration could be very informative. We need to find a tractable way to extract such features.

This study is a first investigation in the new direction of learning to set system configurations. Many underlying questions remain to be addressed in the future. For example, should the learning objective be different from those used in the L2R algorithms? How can we define a different learning algorithm specifically for ranking configurations?

5. REFERENCES

[1] A. Baccini, S. Déjean, L. Lafage, and J. Mothe. How many performance measures to evaluate information retrieval systems? *Knowledge and Information Systems*, 30(3), 2012.

[2] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct. 2001.

[3] D. Carmel and E. Yom-Tov. Estimating the query difficulty for information retrieval. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, 2(1):1–89, 2010.

[4] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29:1189–1232, 2000.

[5] C. Hauff, D. Hiemstra, and F. de Jong. A survey of pre-retrieval query performance predictors. In *Proc. of CIKM*, 2008.

[6] T. Joachims. Optimizing search engines using clickthrough data. In *Proc. of KDD*, 2002.

[7] T.-Y. Liu. Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3(3):225–331, 2009.

[8] C. Macdonald, R. L. Santos, I. Ounis, and B. He. About learning models with multiple query-dependent features. *ACM Transations on Information Systems*, 31(3), 2013.

[9] J. Mothe and L. Tanguy. Linguistic features to predict query difficulty. In *Proc. of SIGIR*, 2005.

[10] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, and C. Lioma. Terrier: A High Performance and Scalable Information Retrieval Platform. In *Proc. of OSIR*, 2006.

[11] A. Shtok, O. Kurland, D. Carmel, F. Raiber, and G. Markovits. Predicting query performance by query-drift estimation. *ACM Transactions on Information Systems*, 30(2):11:1–11:35, 2012.

[12] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.

[13] Y. Zhou and W. B. Croft. Query performance prediction in web search environments. In *Proc. of SIGIR*, 2007.

Table 3: Results with different L2R models and feature abalations. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates figure Statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP		P@100		RPrec	
BM25	0.1942		0.1719		0.2330	
Grid Search	0.2480		0.2213		0.2835	
Random Forests [Opt. nDCG@1] (All)	0.3323 $^{\triangle}$		0.2963 $^{\triangle}$		0.3561 $^{\triangle}$	
Random Forests [Opt. nDCG@10] (All)	0.3274 $^{\triangle}$		0.2851 $^{\triangle}$		0.3500 $^{\triangle}$	
- QueryLing	0.3318 $^{\triangle}$	(+1.34%)	0.2887 $^{\triangle}$	(+1.26%)	0.3497 $^{\triangle}$	(-0.09%)
- QueryStats	0.3321 $^{\triangle}$	(+1.44%)	0.2931 $^{\triangle}$	(+2.81%)	0.3499 $^{\triangle}$	(-0.03%)
- RetModel	0.3016 $^{\triangle}$	(-7.88%)	0.2837 $^{\triangle}$	(-0.49%)	0.3488 $^{\triangle}$	(-0.34%)
- Expansion	0.2478$^{\blacktriangledown}$	(-24.31%)	0.1938$^{\blacktriangledown}$	(-32.02%)	0.2664$^{\blacktriangledown}$	(-23.89%)
- QueryAll	0.3162 $^{\triangle}$	(-3.42%)	0.2731 $^{\triangle}$	(-4.21%)	0.3493 $^{\triangle}$	(-0.20%)
- QueryFeedback	0.3368 $^{\triangle}$	(+2.87%)	0.2749 $^{\triangle}$	(-3.58%)	0.3582 $^{\triangle}$	(+2.34%)
SVM^{rank} (All)	0.2871 $^{\triangle}$		0.2719 $^{\triangle}$		0.3339 $^{\triangle}$	
- QueryLing	0.3306$^{\blacktriangle\triangle}$	(+15.15%)	0.2659 $^{\triangle}$	(-2.21%)	0.3474 $^{\triangle}$	(+4.04%)
- QueryStats	0.3130 $^{\triangle}$	(+9.02%)	0.2540 $^{\triangle}$	(-6.58%)	0.3335 $^{\triangle}$	(-0.12%)
- RetModel	0.2983 $^{\triangle}$	(+3.90%)	0.2590 $^{\triangle}$	(-4.74%)	0.3452 $^{\triangle}$	(+3.38%)
- Expansion	0.1557$^{\blacktriangledown\triangledown}$	(-45.77%)	0.1331$^{\blacktriangledown\triangledown}$	(-51.05%)	0.1877$^{\blacktriangledown\triangledown}$	(-43.79%)
- QueryAll	0.2920 $^{\triangle}$	(+1.71%)	0.2666 $^{\triangle}$	(-1.95%)	0.3615 $^{\triangle}$	(+8.27%)
- QueryFeedback	0.2979 $^{\triangle}$	(+3.76%)	0.2443	(-10.15%)	0.3421 $^{\triangle}$	(+2.46%)
GBRT [Opt. nDCG@1] (All)	0.3289 $^{\triangle}$		0.2772 $^{\triangle}$		0.3411 $^{\triangle}$	
GBRT [Opt. nDCG@10] (All)	0.3442 $^{\triangle}$		0.2586 $^{\triangle}$		0.3176	
- QueryLing	0.3158 $^{\triangle}$	(-8.25%)	0.2644 $^{\triangle}$	(+2.24%)	0.3318 $^{\triangle}$	(+4.47%)
- QueryStats	0.3022$^{\blacktriangledown\triangle}$	(-12.20%)	0.2542 $^{\triangle}$	(-1.70%)	0.3511 $^{\triangle}$	(+10.55%)
- RetModel	0.3099$^{\blacktriangledown\triangle}$	(-9.97%)	0.2603 $^{\triangle}$	(+0.66%)	0.3353 $^{\triangle}$	(+5.57%)
- Expansion	0.2336$^{\blacktriangledown}$	(-32.13%)	0.2064$^{\blacktriangledown}$	(-20.19%)	0.2522$^{\blacktriangledown}$	(-20.59%)
- QueryAll	0.3089$^{\blacktriangledown\triangle}$	(-10.26%)	0.2705 $^{\triangle}$	(+4.60%)	0.2898	(-8.75%)
- QueryFeedback	0.3217 $^{\triangle}$	(-6.54%)	0.2698 $^{\triangle}$	(+4.33%)	0.3246 $^{\triangle}$	(+2.20%)
LambdaMART [Opt. nDCG@1] (All)	0.3074 $^{\triangle}$		0.2655 $^{\triangle}$		0.3196	
LambdaMART [Opt. nDCG@10] (All)	0.3293 $^{\triangle}$		0.2384		0.2960	
- QueryLing	0.2844$^{\blacktriangledown\triangle}$	(-13.63%)	0.2636 $^{\triangle}$	(+10.57%)	0.3210 $^{\triangle}$	(+8.45%)
- QueryStats	0.3110 $^{\triangle}$	(-5.56%)	0.2520	(+5.70%)	0.3294 $^{\triangle}$	(+11.28%)
- RetModel	0.3081 $^{\triangle}$	(-6.44%)	0.2615 $^{\triangle}$	(+9.69%)	0.3352 $^{\triangle}$	(+13.24%)
- Expansion	0.1901$^{\blacktriangledown\triangledown}$	(-42.27%)	0.1878$^{\blacktriangledown}$	(-21.22%)	0.2262$^{\blacktriangledown\triangledown}$	(-23.58%)
- QueryAll	0.2902$^{\blacktriangledown\triangle}$	(-11.87%)	0.2531 $^{\triangle}$	(+6.17%)	0.3135	(+5.91%)
- QueryFeedback	0.2991$^{\blacktriangledown\triangle}$	(-9.17%)	0.2709$^{\blacktriangle\triangle}$	(+13.63%)	0.3283 $^{\triangle}$	(+10.91%)
Upper bound (oracle performance)	0.4136		0.3434		0.4490	

Table 4: Results with different L2R models and feature abalations. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates figure Statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP		P@100		RPrec	
BM25	0.1942		0.1719		0.2330	
Grid Search	0.2617		0.2426		0.2985	
Random Forests [Opt. nDCG@10] (All)	0.3418 $^{\triangle}$		0.2700 $^{\triangle}$		0.3592 $^{\triangle}$	
- QueryLing	0.3259$^{\blacktriangledown\triangle}$	(-4.65%)	0.2686 $^{\triangle}$	(-0.52%)	0.3544 $^{\triangle}$	(-1.34%)
- QueryStats	0.3230 $^{\triangle}$	(-5.50%)	0.2851 $^{\triangle}$	(+5.59%)	0.3404 $^{\triangle}$	(-5.23%)
- RetModel	0.3223 $^{\triangle}$	(-5.71%)	0.2708 $^{\triangle}$	(+0.30%)	0.3409 $^{\triangle}$	(-5.09%)
- Expansion	0.1725$^{\blacktriangledown\triangledown}$	(-49.53%)	0.1902$^{\blacktriangledown\triangledown}$	(-29.56%)	0.2372$^{\blacktriangledown\triangledown}$	(-33.96%)
- QueryAll	0.3232 $^{\triangle}$	(-5.44%)	0.2631	(-2.56%)	0.3727 $^{\triangle}$	(+3.76%)
SVM^{rank} (All)	0.3170 $^{\triangle}$		0.2666		0.3325 $^{\triangle}$	
- QueryLing	0.3206 $^{\triangle}$	(+1.14%)	0.2579	(-3.26%)	0.3426 $^{\triangle}$	(+3.04%)
- QueryStats	0.3018 $^{\triangle}$	(-4.79%)	0.2726 $^{\triangle}$	(+2.25%)	0.3287	(-1.14%)
- RetModel	0.3155 $^{\triangle}$	(-0.47%)	0.2579	(-3.26%)	0.3488 $^{\triangle}$	(+4.90%)
- Expansion	0.1539$^{\blacktriangledown\triangledown}$	(-51.45%)	0.1321$^{\blacktriangledown\triangledown}$	(-50.45%)	0.1890$^{\blacktriangledown\triangledown}$	(-43.16%)
- QueryAll	0.3103 $^{\triangle}$	(-2.11%)	0.2772 $^{\triangle}$	(+3.98%)	0.3366 $^{\triangle}$	(+1.23%)
GBRT [Opt. nDCG@10] (All)	0.3262 $^{\triangle}$		0.2602		0.3393 $^{\triangle}$	
- QueryLing	0.3200 $^{\triangle}$	(-1.90%)	0.2724 $^{\triangle}$	(+4.69%)	0.3346 $^{\triangle}$	(-1.39%)
- QueryStats	0.3256 $^{\triangle}$	(-0.18%)	0.2712 $^{\triangle}$	(+4.23%)	0.3401 $^{\triangle}$	(+0.24%)
- RetModel	0.2952 $^{\triangle}$	(-9.50%)	0.2452	(-5.76%)	0.3313 $^{\triangle}$	(-2.36%)
- Expansion	0.1877$^{\blacktriangledown\triangledown}$	(-42.46%)	0.2000$^{\blacktriangledown\triangledown}$	(-23.14%)	0.2257$^{\blacktriangledown\triangledown}$	(-33.48%)
- QueryAll	0.3275 $^{\triangle}$	(+0.40%)	0.2705 $^{\triangle}$	(+3.96%)	0.3657 $^{\triangle}$	(+7.78%)
LambdaMART [Opt. nDCG@10] (All)	0.2931		0.2472		0.2145 $^{\triangle}$	
- QueryLing	0.3351$^{\blacktriangle\triangle}$	(+14.33%)	0.2579	(+4.33%)	0.2901$^{\blacktriangle}$	(+35.24%)
- QueryStats	0.3119 $^{\triangle}$	(+6.41%)	0.2611	(+5.62%)	0.2525 $^{\triangledown}$	(+17.72%)
- RetModel	0.3126 $^{\triangle}$	(+6.65%)	0.2638 $^{\triangle}$	(+6.72%)	0.3348$^{\blacktriangle\triangle}$	(+56.08%)
- Expansion	0.1747$^{\blacktriangledown\triangledown}$	(-40.40%)	0.1801$^{\blacktriangledown\triangledown}$	(-27.14%)	0.2189 $^{\triangledown}$	(+2.05%)
- QueryAll	0.3211 $^{\triangle}$	(+9.55%)	0.2577	(+4.25%)	0.3175$^{\blacktriangle}$	(+48.02%)
Upper bound (oracle performance)	0.4150		0.3442		0.4490	

Table 5: Results with different L2R models and feature abalations. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates figure Statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP		P@100		RPrec	
BM25	0.1981		0.1628		0.2417	
Grid Search	0.2456		0.1727		0.2725	
Random Forests [Opt. nDCG@10] (All)	0.3067 $^{\triangle}$		0.2015 $^{\triangle}$		0.3205 $^{\triangle}$	
- QueryLing	0.3064 $^{\triangle}$	(-0.10%)	0.1926	(-4.42%)	0.3236 $^{\triangle}$	(+0.97%)
- QueryStats	0.3022 $^{\triangle}$	(-1.47%)	0.2019 $^{\triangle}$	(+0.20%)	0.3173 $^{\triangle}$	(-1.00%)
- RetModel	0.2956 $^{\triangle}$	(-3.62%)	0.1569$^{\blacktriangledown}$	(-22.13%)	0.2720$^{\blacktriangledown}$	(-15.13%)
- Expansion	0.1440$^{\blacktriangledown\triangledown}$	(-53.05%)	0.0969$^{\blacktriangledown\triangledown}$	(-51.91%)	0.2287$^{\blacktriangledown}$	(-28.64%)
- QueryAll	0.3064 $^{\triangle}$	(-0.10%)	0.1929 $^{\triangle}$	(-4.27%)	0.3249 $^{\triangle}$	(+1.37%)
SVM^{rank} (All)	0.2363		0.1655		0.2942	
- QueryLing	0.2426	(+2.67%)	0.1737	(+4.95%)	0.3125 $^{\triangle}$	(+6.22%)
- QueryStats	0.2318	(-1.90%)	0.1689	(+2.05%)	0.3137 $^{\triangle}$	(+6.63%)
- RetModel	0.2523	(+6.77%)	0.1775	(+7.25%)	0.2849	(-3.16%)
- Expansion	0.1383$^{\blacktriangledown\triangledown}$	(-41.47%)	0.1024$^{\blacktriangledown\triangledown}$	(-38.13%)	0.2018$^{\blacktriangledown\triangledown}$	(-31.41%)
- QueryAll	0.2495	(+5.59%)	0.1829	(+10.51%)	0.3002	(+2.04%)
GBRT [Opt. nDCG@10] (All)	0.2585		0.1691		0.2661	
- QueryLing	0.2770	(+7.16%)	0.1643	(-2.84%)	0.2892	(+8.68%)
- QueryStats	0.2749	(+6.34%)	0.1818	(+7.51%)	0.2503	(-5.94%)
- RetModel	0.2438	(-5.69%)	0.1640	(-3.02%)	0.2219 $^{\triangledown}$	(-16.61%)
- Expansion	0.1258$^{\blacktriangledown\triangledown}$	(-51.33%)	0.1025$^{\blacktriangledown\triangledown}$	(-39.38%)	0.2507	(-5.79%)
- QueryAll	0.3013$^{\blacktriangle\triangle}$	(+16.56%)	0.1870	(+10.59%)	0.2987	(+12.25%)
LambdaMART [Opt. nDCG@10] (All)	0.2288		0.1933 $^{\triangle}$		0.2204 $^{\triangle}$	
- QueryLing	0.2523	(+10.27%)	0.1800	(-6.88%)	0.2109 $^{\triangledown}$	(-4.31%)
- QueryStats	0.2027 $^{\triangledown}$	(-11.41%)	0.1879	(-2.79%)	0.2236 $^{\triangledown}$	(+1.45%)
- RetModel	0.2651	(+15.87%)	0.1347$^{\blacktriangledown\triangledown}$	(-30.32%)	0.2212 $^{\triangledown}$	(+0.36%)
- Expansion	0.1307$^{\blacktriangledown\triangledown}$	(-42.88%)	0.0773$^{\blacktriangledown\triangledown}$	(-60.01%)	0.2306	(+4.63%)
- QueryAll	0.2072 $^{\triangledown}$	(-9.44%)	0.1975 $^{\triangle}$	(+2.17%)	0.2272	(+3.09%)
Upper bound (oracle performance)	0.4127		0.2544		0.4493	

Table 6: Robust collection, with optimisation at nDCG@1. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates figure Statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP		P@100		RPrec	
BM25	0.1942		0.1719		0.2330	
Grid Search	0.2617		0.2426		0.2985	
Random Forests [Opt. nDCG@1] (All)	0.3384 △		0.2734 △		0.3558 △	
- QueryLing	0.3312 △	(-2.13%)	0.2709 △	(-0.91%)	0.3559 △	(+0.03%)
- QueryStats	0.3292 △	(-2.72%)	0.2777 △	(+1.57%)	0.3456 △	(-2.87%)
- RetModel	0.3353 △	(-0.92%)	0.2680 △	(-1.98%)	0.3516 △	(-1.18%)
- Expansion	0.1663▼▽	(-50.86%)	0.1923▼▽	(-29.66%)	0.2690▼	(-24.40%)
- QueryAll	0.3262 △	(-3.61%)	0.2598	(-4.97%)	0.3689 △	(+3.68%)
SVM^{rank} (All)	0.3170 △		0.2666		0.3325 △	
- QueryLing	0.3206 △	(+1.14%)	0.2579	(-3.26%)	0.3426 △	(+3.04%)
- QueryStats	0.3018 △	(-4.79%)	0.2726 △	(+2.25%)	0.3287	(-1.14%)
- RetModel	0.3155 △	(-0.47%)	0.2579	(-3.26%)	0.3488 △	(+4.90%)
- Expansion	0.1539▼▽	(-51.45%)	0.1321▼▽	(-50.45%)	0.1890▼▽	(-43.16%)
- QueryAll	0.3103 △	(-2.11%)	0.2772 △	(+3.98%)	0.3366 △	(+1.23%)
GBRT [Opt. nDCG@1] (All)	0.3210 △		0.2877 △		0.3440 △	
- QueryLing	0.3140 △	(-2.18%)	0.2858 △	(-0.66%)	0.3491 △	(+1.48%)
- QueryStats	0.3220 △	(+0.31%)	0.2597	(-9.73%)	0.3542 △	(+2.97%)
- RetModel	0.2921	(-9.00%)	0.2679	(-6.88%)	0.3194	(-7.15%)
- Expansion	0.2285▼	(-28.82%)	0.1892▼▽	(-34.24%)	0.2144▼▽	(-37.67%)
- QueryAll	0.3321 △	(+3.46%)	0.2738 △	(-4.83%)	0.3477 △	(+1.08%)
LambdaMART [Opt. nDCG@1] (All)	0.2237		0.2729 △		0.3024	
- QueryLing	0.2725	(+21.81%)	0.2223▼	(-18.54%)	0.2684	(-11.24%)
- QueryStats	0.2402	(+7.38%)	0.2770 △	(+1.50%)	0.3114	(+2.98%)
- RetModel	0.3127▲△	(+39.79%)	0.2442	(-10.52%)	0.3070	(+1.52%)
- Expansion	0.2080 ▽	(-7.02%)	0.1688▼▽	(-38.15%)	0.1920▼▽	(-36.51%)
- QueryAll	0.3226▲△	(+44.21%)	0.2722 △	(-0.26%)	0.3426▲△	(+13.29%)
Upper bound (oracle performance)	0.4150		0.3442		0.4490	

Table 7: WT10g collection, with optimisation at nDCG@1. △ indicates statistically significant improvements over the Grid Search baseline, according to a paired t-test ($p < 0.05$). ▼ indicates figure Statistically significant decreases induced by a feature ablation with respect to the corresponding (All) models.

	MAP		P@100		RPrec	
BM25	0.1981		0.1628		0.2417	
Grid Search	0.2456		0.1727		0.2725	
Random Forests [Opt. nDCG@1] (All)	0.3109 $^{\triangle}$		0.2020 $^{\triangle}$		0.3171 $^{\triangle}$	
- QueryLing	0.2947 $^{\triangle}$	(-5.21%)	0.1955 $^{\triangle}$	(-3.22%)	0.3268 $^{\triangle}$	(+3.06%)
- QueryStats	0.2922 $^{\triangle}$	(-6.01%)	0.1992 $^{\triangle}$	(-1.39%)	0.3280 $^{\triangle}$	(+3.44%)
- RetModel	0.2768$^{\blacktriangledown\triangle}$	(-10.97%)	0.1743$^{\blacktriangledown}$	(-13.71%)	0.2677$^{\blacktriangledown}$	(-15.58%)
- Expansion	0.1443$^{\blacktriangledown\triangledown}$	(-53.59%)	0.1118$^{\blacktriangledown\triangledown}$	(-44.65%)	0.2179$^{\blacktriangledown\triangledown}$	(-31.28%)
- QueryAll	0.3066 $^{\triangle}$	(-1.38%)	0.1999 $^{\triangle}$	(-1.04%)	0.3174 $^{\triangle}$	(+0.09%)
SVM^{rank} (All)	0.2363		0.1655		0.2942	
- QueryLing	0.2426	(+2.67%)	0.1737	(+4.95%)	0.3125 $^{\triangle}$	(+6.22%)
- QueryStats	0.2318	(-1.90%)	0.1689	(+2.05%)	0.3137 $^{\triangle}$	(+6.63%)
- RetModel	0.2523	(+6.77%)	0.1775	(+7.25%)	0.2849	(-3.16%)
- Expansion	0.1383$^{\blacktriangledown\triangledown}$	(-41.47%)	0.1024$^{\blacktriangledown\triangledown}$	(-38.13%)	0.2018$^{\blacktriangledown\triangledown}$	(-31.41%)
- QueryAll	0.2495	(+5.59%)	0.1829	(+10.51%)	0.3002	(+2.04%)
GBRT [Opt. nDCG@1] (All)	0.2485		0.1699		0.2496	
- QueryLing	0.2812 $^{\triangle}$	(+13.16%)	0.1599	(-5.89%)	0.2567	(+2.84%)
- QueryStats	0.2695	(+8.45%)	0.1934 $^{\triangle}$	(+13.83%)	0.2818	(+12.90%)
- RetModel	0.2910$^{\blacktriangle\triangle}$	(+17.10%)	0.1582	(-6.89%)	0.2522	(+1.04%)
- Expansion	0.1678$^{\blacktriangledown\triangledown}$	(-32.47%)	0.1124$^{\blacktriangledown\triangledown}$	(-33.84%)	0.2572	(+3.04%)
- QueryAll	0.2938$^{\blacktriangle\triangle}$	(+18.23%)	0.1819	(+7.06%)	0.3198$^{\blacktriangle\triangle}$	(+28.12%)
LambdaMART [Opt. nDCG@1] (All)	0.2013 $^{\triangle}$		0.1391		0.2285 $^{\triangle}$	
- QueryLing	0.1905 $^{\triangledown}$	(-5.37%)	0.1436	(+3.24%)	0.1734$^{\blacktriangledown\triangledown}$	(-24.11%)
- QueryStats	0.2701$^{\blacktriangle}$	(+34.18%)	0.1694	(+21.78%)	0.2022 $^{\triangledown}$	(-11.51%)
- RetModel	0.0935$^{\blacktriangledown\triangledown}$	(-53.55%)	0.1807$^{\blacktriangle}$	(+29.91%)	0.2027 $^{\triangledown}$	(-11.29%)
- Expansion	0.1592 $^{\triangledown}$	(-20.91%)	0.0880$^{\blacktriangledown\triangledown}$	(-36.74%)	0.2429	(+6.30%)
- QueryAll	0.2123	(+5.46%)	0.1729	(+24.30%)	0.2774	(+21.40%)
Upper bound (oracle performance)	0.4127		0.2544		0.4493	