Exploiting Result Diversification Methods for Feature Selection in Learning to Rank

Kaweh Djafari Naini[1] and Ismail Sengor Altingovde[2]

[1] L3S Research Center, Leibniz University Hannover, Hannover, Germany
naini@l3s.de
[2] Middle East Technical University, Ankara, Turkey
altingovde@ceng.metu.edu.tr

Abstract. In this paper, we adopt various greedy result diversification strategies to the problem of feature selection for learning to rank. Our experimental evaluations using several standard datasets reveal that such diversification methods are quite effective in identifying the feature subsets in comparison to the baselines from the literature.

1 Introduction

Learning to rank (LETOR) is the state of the art method employed by the large-scale commercial search engines to rank the search results. Given the large number of features available in a search engine, which is in the order of several hundreds (e.g., see Yahoo! LETOR Challenge), it is desirable to identify a subset of features that yield a comparable effectiveness to using all the features. Since search engines typically employ a two-stage retrieval where an initial set of candidate documents are re-ranked using a sophisticated LETOR model, a smaller number of features would reduce the feature computation time, which must be done on-the-fly for the query dependent features, and hence overall query processing time. Furthermore, improving the efficiency of the LETOR stage would allow retrieving larger candidate sets and, subsequently, can help enhancing the quality of the search results.

In a recent study, Geng et al. proposed a filtering-based feature selection method that aims to select a subset of features that are both effective and dissimilar to each other [5]. Inspired from this study, we draw an analogy between the feature selection and result diversification problems. In the literature, a rich set of greedy diversification methods are proposed to select both relevant and diverse top-k results for web search queries (e.g., see [1,6,12,9,10]). We apply three representative diversification methods, namely, Maximal Marginal Relevance (MMR) [1], MaxSum Dispersion (MSD) [6] and Modern Portfolio Theory (MPT) [12,9] to the feature selection problem for LETOR. To the best of our knowledge, none of these methods are employed in the context of learning to rank with the standard search engine datasets.

In the next section, we first describe the baseline strategies for the feature selection from the literature, and then discuss how we adopt the result diversification methods for this purpose. In Sections 3 and 4, we present the experimental setup and evaluation results, respectively. Finally, we conclude in Section 5.

M. de Rijke et al. (Eds.): ECIR 2014, LNCS 8416, pp. 455–461, 2014.

2 Feature Selection for LETOR

Feature selection techniques for the classification tasks are heavily investigated in the literature and fall into three different categories, namely, filter, wrapper and embedded approaches [5]. Strategies in the filter category essentially work independently from the classifiers and choose the most promising features in a preprocessing step. In contrast, the strategies following the wrapper approach consider the metric that will be optimized by the classifier whereas those in the embedding category incorporate the feature selection into the learning process. Earlier studies also show that such feature selection methods do not only help improving the accuracy and efficiency of the classifiers, but may also introduce diversity in ensembles of classifiers [3].

For learning to rank, there are only a few recent studies that address the feature selection issue [5,4]. Following the practice in [5], we focus on the feature selection methods that fall into the filter category.

2.1 Preliminaries

For a given feature $f_i \in F$, we obtain its relevance score for a query by ranking the results of a query solely on this feature and computing the effectiveness for the top-10 results. The effectiveness can be measured using any well-known evaluation measure (like MAP, NDCG) or a loss function (as in [5]). In this study, we employ NDCG@10 as the effectiveness measure and denote the *average* relevance score of a feature over all queries by $rel(f_i)$. To capture the similarity of any two features, denoted with $sim(f_i, f_j)$, we compute the Kendall's Tau distance between their top-10 rankings averaged over all queries (as in [5]). The objective is selecting a subset of k features (F_k), where $k < |F|$, such that both the relevance and diversity (dissimilarity) among the selected features are maximized.

2.2 Baseline Feature Selection Methods

Top-k Relevant (TopK): A straightforward method for feature selection is choosing the top-k features that individually yield the highest average relevance scores over the queries [4].

Greedy Search Algorithm (GAS): This is the greedy strategy proposed by Geng et al. in [5]. It starts with choosing the feature, say f_i, with the highest average relevance score into the set F_k. Next, for each of the remaining features f_j, its relevance score is updated with respect to the following equation:

$$rel(f_j) = rel(f_j) - sim(f_i, f_j) \cdot 2c, \tag{1}$$

where c is a parameter to balance the relevance and diversity optimization objectives. The algorithm proceeds in a greedy manner by choosing the next feature with the highest score and updating the remaining scores, until k features are determined.

2.3 Diversification Methods for Feature Selection

As the astute reader would realize, the goal of feature selection as defined in [5,4] is identical to that of the search result diversification techniques: both problems require selecting the most relevant and, at the same time, diverse items. Motivated by this observation, we adopt three different implicit result diversification techniques to the feature selection problem, as follows.

Maximal Marginal Relevance (MMR): This is a well-known greedy strategy originally proposed in [1]. Peng et al. propose a similar idea of minimal-redundancy maximal-relevance in [8]. In a recent study [2], MMR is employed for feature selection in learning to rank in a setup with a limited number of social features, but not evaluated on the standard search datasets, as we do in this paper.

In this study, we adopt a version of MMR described in [11]. The MMR strategy also starts with choosing the feature f_i with the highest relevance score into the F_k. At each iteration, MMR computes the score of an unselected feature f_j according to the following equation:

$$mmr(f_j) = (1-\lambda)rel(f_j) + \frac{\lambda}{|F_k|}\sum_{f_i \in F_k} 1 - sim(f_i, f_j), \lambda \in [0,1], \qquad (2)$$

where λ is again a trade-off parameter to balance the relevance and diversity.

MaxSum Dispersion (MSD): An alternative representation of the diversification (and hence, feature selection) problem is casting it to the facility dispersion problem in the operations research field [6]. In this case, our objective in this paper, i.e., maximizing the sum of relevance and dissimilarity in F_k, can be solved with the greedy 2-approximation algorithm that is originally proposed for the well-known MaxSum Dispersion (MSD) problem. In the MSD solution, a pair of features that maximizes the following equation is selected into F_k at each iteration:

$$msd(f_i, fj) = (1-\lambda)(rel(f_i) + rel(f_j)) + 2\lambda(1 - sim(f_i, f_j)), \qquad (3)$$

where λ is the trade-off parameter.

Modern Portfolio Theory (MPT): This approach is based on the famous financial theory which states that one should diversify her portfolio by maximizing the expected return (i.e, mean) and minimizing the involved risk (i.e., variance). In case of the result diversification, this statement implies that we have to select the documents that maximize the relevance and have a low variance of relevance [12,9]. The latter component has to be treated as a parameter and its best value can be computed by sweeping through the possible values (as in [12]) unless additional data, such as click logs, are available [9].

Fortunately, in case of the feature selection for LETOR, we have adequate data to model both the mean and variance of the relevance of a feature. Obviously, mean relevance of a feature is $rel(f_i)$ as we have already defined. For the variance of a feature ($\sigma^2(f_i)$), we compute the relevance score of f_i for each query q, and then compute the variance for this set of scores in a straightforward manner.

Table 1. Datasets

Dataset	No. of queries	No. of annotated results	No. of features
OHSUMED	106	16,140	45
MQ2008	800	15,212	46
Yahoo! SET2	6,330	172,870	596

Thus, the greedy MPT solution chooses the feature that maximizes the following equation at each iteration:

$$mpt(f_i) = rel(f_i) - [b\sigma^2(rel(f_i)) + 2b\sigma(rel(f_i)) \sum_{f_j \in F_k} \sigma(rel(f_j)) * sim(f_i, f_j)]. \quad (4)$$

Note that, we eliminated the rank position component from the original formula [12,9] as it does not make sense for the feature selection problem. As before, $b \in [0, 1]$ is the trade-off parameter to balance the relevance and diversity.

3 Experimental Setup

Datasets. Our experiments are conducted on three standard LETOR datasets, OHSUMED[1] from Letor3.0, MQ2008 from Letor4.0 and SET2[2] from Yahoo! LETOR Challenge. In Table 1 we summarize the characteristics of each dataset. The Yahoo! SET2 has 596 features and is also the largest dataset with respect to the number of queries and instances. But previous studies have also shown that even for a small number of features, feature selection can improve the ranking [5].

LETOR Algorithm. Our evaluations employ RankSVM [7], which is a very widely used pairwise LETOR algorithm. More specifically, we used SVMRank[3] library implementation. We trained the classifier with a linear kernel with $\epsilon = 0.001$. We report the results with the C values (where $C \in [0.00001, 10]$) that yields the best performance on the test set with all the features.

Evaluation Measures. We evaluate all the feature selection methods using 5-fold cross validation for the OHSUMED and MQ2008 datasets. For Yahoo! SET2, we use the training and test sets as provided. The evaluation measures are MAP and NDCG@10.

4 Experimental Results

In Figure 1, we report the NDCG@10 and MAP scores obtained on the OHSUMED dataset using the baseline and proposed feature selection methods. We observe that when the number of selected features is greater than 10, the performance is comparable or better than using all features (ALL). Furthermore, the methods adapted from the diversity field outperform the baselines (TopK and GAS). In

[1] `http://research.microsoft.com/en-us/um/beijing/projects/letor`
[2] `http://webscope.sandbox.yahoo.com/catalog.php?datatype=c`
[3] `http://www.cs.cornell.edu/people/tj/svm_light/svm_rank.html`

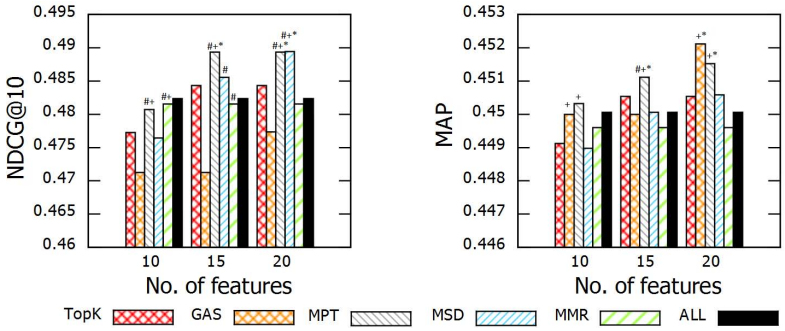

Fig. 1. Ranking effectiveness on OHSUMED: NDCG@10 (left) and MAP (right)

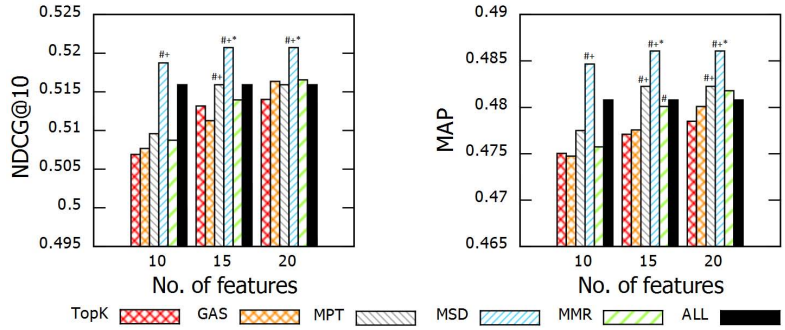

Fig. 2. Ranking effectiveness on MQ2008: NDCG@10 (left) and MAP (right)

particular, MPT is the winner for both evaluation measures when the number of features is set to 15 or 20.

Figure 2 shows the performance for the MQ2008 dataset. In this case, the feature selection algorithms can reach the performance of the ALL only after selecting more than 15 features. For the majority of the cases, the methods adopted from the diversification field are again superior to the baselines, and MSD is the winner method for this dataset.

Finally, in Figure 3, we report the performance for the Yahoo! SET2. As the experiments take much larger time on this dataset, we only present the results for selecting 100 features (out of 596). We observe that, feature selection methods with 100 features cannot beat the all features baseline ALL (not shown in the plots), which is reasonable as we only use one sixth of the available features. MPT is again the best adapted method, and it outperforms TopK baseline for both evaluation measures, and better than or comparable to GAS for MAP and NDCG measures, respectively.

The statistical significance of our methods is verified using the paired t-test with $p < 0.05$. In Figures 1-3, we show the significant differences to the baselines TopK (denoted with +), GAS (denoted with #) and ALL (denoted with *).

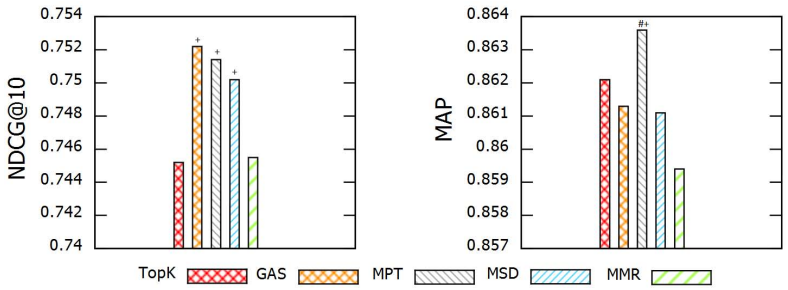

Fig. 3. Ranking effectiveness on the Yahoo! SET2: NDCG@10 (left) and MAP (right)

5 Conclusions

We adopted several methods from the result diversification field to address the problem of feature selection for LETOR. Our evaluations showed that these methods yield higher effectiveness scores than the baseline feature selection strategies for various standard datasets.

Acknowledgments. This work was partially funded by the European Commission FP7 under grant agreement No. 600826 for the ForgetIT project and The Scientific and Technological Research Council of Turkey (TÜBİTAK) under the grant no. 113E065. I. S. Altingovde acknowledges the Yahoo! FREP.

References

1. Carbonell, J.G., Goldstein, J.: The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: Proc. of SIGIR 1998 (1998)
2. Chelaru, S.V., Orellana-Rodriguez, C., Altingovde, I.S.: How useful is social feedback for learning to rank youtube videos? WWW Journal, 1–29 (in press), doi:10.1007/s11280-013-0258-9
3. Cunningham, P., Carney, J.: Diversity versus quality in classification ensembles based on feature selection. In: Lopez de Mantaras, R., Plaza, E. (eds.) ECML 2000. LNCS (LNAI), vol. 1810, pp. 109–116. Springer, Heidelberg (2000)
4. Dang, V., Croft, W.B.: Feature selection for document ranking using best first search and coordinate ascent. In: Proc. SIGIR 2010 Workshop on Feature Generation and Selection for Information Retrieval (2010)
5. Geng, X., Liu, T.-Y., Qin, T., Li, H.: Feature selection for ranking. In: Proc. of SIGIR 2007, pp. 407–414 (2007)
6. Gollapudi, S., Sharma, A.: An axiomatic approach for result diversification. In: Proc. of WWW 2009, pp. 381–390 (2009)
7. Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers, pp. 115–132 (2000)
8. Peng, H., Long, F., Ding, C.H.Q.: Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans. Pattern Anal. Mach. Intell. 27(8), 1226–1238 (2005)

9. Rafiei, D., Bharat, K., Shukla, A.: Diversifying web search results. In: Proc. of WWW 2010, pp. 781–790 (2010)
10. Santos, R.L.T., Castells, P., Altingovde, I.S., Can, F.: Diversity and novelty in information retrieval. In: Proc. of SIGIR 2013, p. 1130 (2013)
11. Vieira, M.R., Razente, H.L., Barioni, M.C.N., Hadjieleftheriou, M., Srivastava, D., Train Jr., C., Tsotras, V.J.: On query result diversification. In: Proc. of ICDE 2011, pp. 1163–1174 (2011)
12. Wang, J., Zhu, J.: Portfolio theory of information retrieval. In: Proc. of SIGIR, pp. 115–122 (2009)