# Evolutionary Risk-Sensitive Feature Selection for Learning to Rank

**Daniel Xavier de Sousa (author)**[1,2], **Thierson Couto Rosa (co-advisor)**[3], **Marcos André Gonçalves(advisor)**[2]

[1]Instituto Federal de Goiás (IFG)
Anápolis – GO – Brazil

[2]Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brazil

[3]Instituto de Informática – Universidade Federal de Goiás (UFG)
Goiânia – GO – Brazil

daniel.sousa@ifg.edu.br,mgoncalv@dcc.ufmg.br, thierson@inf.ufg.br

## 1. Introduction

Learning to Rank (L2R) has established itself as an important research area in Information Retrieval (IR). This is because L2R is the central task in many important IR applications such as modern Web search engines, recommendation and question-answering systems [Chapelle et al. 2011]. In general, L2R applies machine learning algorithms to improve the ranking quality by using annotated information about the relevance of documents.

To obtain good results, L2R strategies usually rely on dense representations exploiting dozens of features, some of which are expensive to generate. In several scenarios, some of these features may introduce noise or may be redundant, increasing the cost of the learning process without bringing benefits or even harming the learned ranking model.

Thus, Feature Selection (FS) techniques have been examined in the L2R scenario [Laporte et al. 2014] to improve processing time and increase effectiveness by removing noisy and redundant features. FS indeed may have a high positive impact on processing time in L2R [Chapelle et al. 2011]. In addition to the training time, there is also the cost of constructing the features (actually meta-features) as they are generated by several algorithms (e.g., BM25, PageRank) and some of them need to be computed at query time.

Nevertheless, effectiveness and cost (better summarized by the number of exploited features) are not the only objectives one may want to optimize in a L2R task. In fact, recently the **risk** of getting very poor effectiveness for a few queries with a learned model has gained much attention [Dinçer et al. 2016]. This interest in diminishing risk is due mainly to the fact that users tend to remember the few failures of a search engine very well rather than the many successful searches. In fact, the authors in [Zhang et al. 2014] clearly show that improvements in ranking performance do not always correlate with risk reduction. This has motivated research in *risk-sensitive L2R* which considers the risk aspect of L2R models. The goal of the risk-sensitive L2R task is to enhance the overall effectiveness of a ranking system while reducing the risk of performing poorer than a baseline ranking system for any given query.

Therefore, in our dissertation we claim that feature selection used with the intent specifically to enhance efficiency and effectiveness may be a problem to risk-sensitiveness

in L2R. This happens because FS reduces the feature space when considering only overall effectiveness or cost as objectives. It is possible that the reduction of features may worsen the ranking of documents for a few queries (but important ones, such as medical searches), despite improving the ranking for many others. Thus, there may be features that, despite not significantly improving the ranking effectiveness average, enhance the quality of few queries, providing a more robust performance.

Figure 1 provides evidence of the above claim, which shows in $x$-axis different rankings (one for each column) when using the effectiveness and risk-sensitive measure to sort the features of the MLSR-WEB10K dataset[1]. The first ranking sorts the features considering effectiveness, measured in terms of NDCG@10. The other four rankings correspond to the same features using four different weights of the GeoRisk risk-sensitive function[2]. Each feature corresponds to a colored line in the figure, guided by the rank position of features (in $y$-axis) in each ranking.

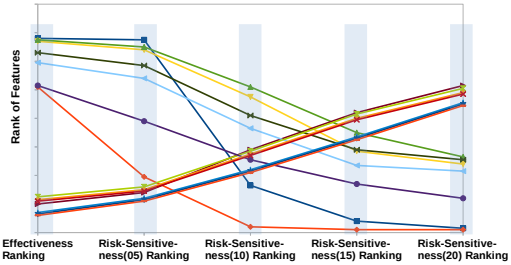


**Figura 1. Ranking features when varying the measures of sorting.**

Figure 1 shows that some features have an essential behavior on ranking effectiveness, but they are less important from a risk-sensitive perspective, whereas the opposite occurs with other features. In other words, Figure 1 illustrates an essential aspect of FS in L2R: the filtering of features considering only the optimization of effectiveness as a criterion may prune important features that would help to generate more robust (less risky) models. Hence, the problem proposed in this dissertation concerns the selection of features with risk-sensitiveness as a main objective criterion, without loss of effectiveness. Furthermore, as described in our dissertation, the selection of features when using effectiveness as a single objective criterion may incur in higher risk, mainly because the methods tend to optimize an average metric such as Mean Average Precision (MAP) or NDCG, despite potential losses in few points.

## 1.1. Research Goals

The above observations have motivated us to address distinct objective criteria in FS for the L2R task. To the best of our knowledge, there are no studies that provide a thorough analysis of the impact of feature selection in both effectiveness and risk-sensitiveness in the L2R literature. Accordingly, the novel proposed objective criteria are: i) maximizing the ranking effectiveness; ii) minimizing the risk for most queries; and iii) reducing the feature space dimensionality, *all at the same time*. Moreover, we analyze the impact of

[1]MLSR-WEB10K is a public dataset, released by Microsoft with 10,000 queries and 136 features.

[2]GeoRisk provides a risk-sensitive evaluation of model performance, by comparing against a set of baselines. The weights ponder the degradation effect or negative variation of the evaluated model against a set of baselines.

FS for L2R in these objectives considering them both individually (as single objectives), as well as combined as multi-objectives to be optimized.

By considering a robust and effective evaluation with FS, our dissertation aims to obtain a possibly smaller set of features that guarantees ranking effective and risk-sensitive performance. This is in contrast to existing FS for L2R approaches which goal is to drastically reduce the number of features in order to control the processing time.

We also propose a novel methodology to assess the impact on effectiveness and risk-sensitiveness when diverse (most of the times, conflicting) objective criteria are applied to the FS for the L2R task. Using an efficient and effective wrapper strategy, our proposed methodology explores diverse sets of features as a search space and uses an evolutionary search to select the best features set according to single or multi-objective criteria. Wrapper strategies are traditionally recognized as time consuming approaches [Laporte et al. 2014]. To deal with this issue, in our dissertation we propose to exploit "cheap" weak-learners as black-boxes that make the process more scalable and less costly. As a positive side effect, weak-learners also promote diversity in the solutions, as strong-learners are more agnostic to the employed set of features. In other words, we drive the exploration of the space of solutions using improvements in both wrapper and multi-objective optimization processing.

To investigate combinations of simultaneous objectives, our proposal uses a multi-objective criteria approach based on Pareto frontier optimization. There are several general-purpose multi-objective optimization methods that can be used in this case. We have chosen Strength Pareto Evolutionary Algorithm (SPEA2) [Zitzler et al. 2001], which besides being the state-of-the-art in multi-objective optimization, has already been successfully applied to several related problems [Li et al. 2015]. In fact, Evolutionary Algorithms (EAs) are well suited to estimate the impact of the distinct proposed objective criteria and also to evaluate our statements, mainly due to their capability of obtaining non-linear ranking functions. In summary, in our dissertation we provide three novel contributions:

1. We open up a new perspective of FS for L2R, which highlights the importance of considering risk as an explicit objective criterion. In this context, we are not only considering the average effectiveness obtained by a drastically reduced subset of features, but a subset which provides a risk-sensitive and effective performance;
2. We introduce single and multi-objective criteria to perform FS for L2R, considering three important objectives, *concomitantly*: feature dimensionality reduction, effectiveness and risk-sensitiveness. Some of these (conflicting) objective criteria were never evaluated in FS for L2R;
3. A novel efficient and effective evolutionary methodology to evaluate different objective criteria in FS for L2R. We apply weak-learners (apparently counter intuitive) to decrease the execution time while increasing diversity, and a paired test comparison over a multi-objective search to provide an accurate set of features.
4. We provide a broad discussion of the proposed methodology and objectives, showing that, in FS for L2R, distinct goals (with feature reduction or accuracy) can be achieved by varying the objective criteria. Also, most previous works explored only small datasets, and we consider large ones, e.g. WEB10K and YAHOO.

During the dissertation, we published some papers in the world leading In-

formation Retrieval conferences and journals, such as [Sousa et al. 2016]*(A1)* and [Sousa et al. 2019]*(A2)*. Beside them, we also published other papers in L2R area, as [Sousa et al. 2012]*(B1)*, [Freitas et al. 2016]*(B3)*[3], and [Freitas et al. 2018]*(A2)*.

## 2. Experimental Evaluations

As our work explores a new direction in FS for L2R, several interesting experimental evaluation were performed, as we detail in the following.

The computation of the fitness value for an individual in a wrapper strategy is time consuming, as it is necessary to construct a hard L2R model with a subset of features corresponding to the individual and to evaluate this model to derive the values for effectiveness and risk-sensitive measures. This has to be done for each individual in the population and is specially time consuming for some large datasets and state-of-the-art L2R algorithms. Hence, one of the key points in our work is the reduction of the searching time during the wrapper-based feature selection, by applying a weak-leaner as a cost-function to evaluate the individuals over the evolutionary search.

In addition, the literature shows that the Pareto frontier set can be large, especially when two objectives are conflicting. This can make the selection within the Pareto set very hard, decreasing the final performance. We address this selection using a strict comparison over the individuals by the mean of an evolutionary search, using statistical hypothesis tests. As a result, our method provides a smaller Pareto set with only statistically superior individuals. This has an important impact on the accuracy of our methods, as is described in our results.

Moreover, differently from all other works in the literature of FS for L2R, we here describe the performance of many objective criteria from a risk-sensitive perspective, and we show that risk-sensitiveness is an important objective criterion in FS. We also provide a full evaluation of many objective criteria over three dimensions, *concomitantly*: ranking performance, dimensionality, and risk-sensitiveness. Considering the several intents over FS, we provide clear demonstrations of results for the objectives and evaluated datasets.

Even though we combine multi-objective criteria and wrapper strategy to find a better feature interaction to build a model, we believe that it is possible to evaluate features which improve the risk-sensitiveness rather than effectiveness. In other words, we drive our attention to point out which specific features or group of features provide more impact on risk-sensitiveness when building the L2R model. We show that there are some features which despite not being applied to optimize the effectiveness criterion, are used to support the robustness for some other queries.

To conclude, as we have proposed distinct strategies to improve risk-sensitiveness and effectiveness for FS in L2R, we have paid attention to the effect of each proposal in the experimental results, performing a $2^k$ Factorial Design to discover the result variation obtained for each measure, i.e. risk-sensitiveness and effectiveness. As a result, we have observed that the statistical test also improves the risk-sensitiveness, as it performs a model comparison concerning all available queries.

[3]Selected as the best paper of the conference

## 3. Conclusion

Our work is the first dissertation that thoroughly investigated the impact of risk-sensitiveness in feature selection for Learning to Rank. In this context, we provided relevant conclusions described in next sections.

### 3.1. A New Methodology to Evolutionary Algorithms

In our dissertation we perform a multi-objective criteria using SPEA2 as a general multiobjective criteria, concerning the interaction of features on the wrapper strategy and without being attached to a particular L2R algorithm as a black-box. We noted that this strategy provides the flexibility to search for several regions in the feature space, even providing feature reduction without using number of feature as a objective criterion.

Beside that, our methodology extends the evolutionary wrapper algorithms by using weak learners as black-box. We show that weak learners, e.g. Linear Regression and Regression Tree, can be applied in a wrapper strategy to perform FS on the L2R task and by improving the time cost more than 120x, without decreasing the effectiveness. Furthermore, we note that a weak learner allows a more sensitive comparison to evaluate the individuals, as it assigns the fitness values penalizing the presence of bad features. In contrast to strong-learners, which can decrease the weight of bad features in time to build the model, providing similar effectiveness when comparing distinct sets of features. After a set of features is selected, a state-of-the-art L2R algorithm is used as the final model.

Our methodology also extends the works in Pareto set, as it applies strict comparison among individuals, by performing a paired statistical test to define the dominance relationship of the individual over the generations. As described in our experiments, this strategy reduces the conflict between individuals in multi-objective criteria, decreasing the Pareto set dimension over 50% and improving the effectiveness of the selected individual.

### 3.2. Risk-sensitive Feature Selection for Learning to Rank

In our dissertation we stress the evaluation of several risk-sensitive measures with effectiveness and number of features as multi-objective to perform feature selection. As a result, we show that using effectiveness and risk-sensitiveness as objective criteria provide a better subset of features for L2R, which increases the effectiveness of most queries and also of some queries which are avoided in the absence of risk-sensitive measures. Moreover, we observe that using a combination of multi-objective criteria is better than using a single one, even when the main goal is used as the objective.

In fact, we show that risk-sensitiveness is an important objective criterion in order to perform FS in L2R. In our experimental results, all methods which have only effectiveness and/or number of features as an objective criterion could not outperform the methods which have risk-sensitiveness and effectiveness as criteria. This is an important contribution in the area, as effectiveness and feature reduction are commonly found in works of literature for FS in L2R.

Our work improves the feature-space by i) reducing the time performance to execute the L2R phases, making it more flexible to update the training set; ii) evaluating several feature selection objectives, for instance, reducing the feature dimensionality without damaging effectiveness; and mainly, iii) with the responsibility of not increasing the risk of obtaining bad predictions for some queries.

We also evaluate the quality of specific features over effectiveness and risk-sensitiveness. We may say that there are groups of features which are not important to improve the overall effectiveness, such as features *"Sum"* and *"Min of stream length normalized term frequency"* in WEB10K. However, as these features are important to improve the effectiveness of same queries, they are selected in case of risk-sensitiveness as a base criterion. This is a very interesting result, as some features can now be considered important for the risk-sensitive perspective.

In addition, we observe that our proposal have distinct impact in results variation, performing a $2^k$ Factorial Design. For instance, we observe that the statistical test also improves the risk-sensitive performance when selection the feature, allowing a robust comparison over all queries. Moreover, our experimental results describes that the risk-sensitiveness as a objective-criteria also improves the effective performance.

To sum up, from our experimental results we note that besides the features provide distinct important roles in the feature-space, e.g. low-risk and/or effectiveness, the rate of unimportant features (noisy and redundant) is absolutely uncertain in datasets, with some datasets having more of this kind of features than others. Hence, the task of selecting relevant set of features becomes even more challenging. In this sense, our dissertation provides a relevant and a novel contribution in FS for L2R, by including the risk-sensitiveness as a criterion in FS, enhancing the selection of Pareto set individuals and the processing time of wrapper strategies.

## Referências


Chapelle, O., Yi, C., and Liu, T.-Y. (2011). Future directions in learning to rank. *In YLRC*, pages 129–136.

Dinçer, B. T., Macdonald, C., and Ounis, I. (2016). Risk-Sensitive Evaluation and Learning to Rank using Multiple Baselines. *In SIGIR*, pages 483–492.

Freitas, M., Sousa, D., Martins, W., Couto, T., Silva, R., and Gonçalves, M. (2016). A Fast and Scalable Manycore Implementation for an On-Demand Learning to Rank Method. *In WSCAD*, 1:1–12.

Freitas, M., Sousa, D., Martins, W., Couto, T., Silva, R., and Gonçalves, M. (2018). Parallel rule-based selective sampling and on-demand learning to rank. *In CCPE*, pages 1–12.

Laporte, L., Flamary, R., Canu, S., Dejean, S., and Mothe, J. (2014). Nonconvex regularizations for feature selection in ranking with sparse SVM. *In IEEE TNNLS*, abs/1507.00500:1118–1130.

Li, B., Li, J., and Tang, K. (2015). Many-Objective Evolutionary Algorithms: A Survey. *In CSUR*, 48:1–35.

Sousa, D., Canuto, S., Couto, T., Martins, W., and Gonçalves, M. (2016). Incorporating Risk-Sensitiveness into Feature Selection for Learning to Rank. *In CIKM*, pages 257–266.

Sousa, D., Canuto, S., Gonçalves, M. A., Couto, T., and Martins, W. (2019). Risk-sensitive learning to rank with evolutionary multi-objective feature selection. *In TOIS*, 37:24:1–24:34.

Sousa, D., Couto, T., Martins, W., Silva, R., and Gonçalves, M. (2012). Improving on-demand learning to rank through parallelism. *In WISE*, pages 526–537.

Zhang, P., Hao, L., Song, D., Wang, J., Hou, Y., and Hu, B. (2014). Generalized Bias-Variance Evaluation of TREC Participated Systems. *In CIKM*, pages 3–6.

Zitzler, E., Laumanns, M., and Thiele, L. (2001). SPEA2: Improving the strength pareto evolutionary algorithm. *In EUROGEN*, pages 12–19.