In the context of supervised learning, feature selection is the task of selecting a subset of features that is most useful for the prediction task at hand. In [1], a feature selection method for the problem of ranking has been proposed. Here, feature selection is formalized as an optimization task that seeks to find a compromise between two criteria: maximizing the total importance of the selected features while minimizing the total similarity between them. The MAP and NDCG measures (widely used measures in information retrieval) are used for computing the importance of a feature, and Kendall’s $\tau$ as a metric for similarity between features.

In spite of the usefulness of the proposed technique, there seems to be scope for potential improvements: (i) Instead of combining the two objectives (importance and dissimilarity) linearly, as done in the paper, one could think of other combinations. (ii) Likewise, the choice of measures such as MAP, NDCG, and Kendall’s $\tau$ appears to be arbitrary to some extent, and perhaps other measures could improve performance. (iii) While importance and similarity of the features are considered in [1], one could go one step further and look at other objectives as well, such as redundancy of features.

The goal of the thesis is to elaborate on variations and extension of the feature selection method of [1], to implement these variations, and to evaluate them experimentally.

Requirements: Design of algorithms for feature selection in ranking; implementation of the algorithms; collection of suitable data for experimental studies; experimental evaluation and analysis of results.

Prerequisites: Basic knowledge in machine learning and data analysis, programming skills.

Contact: Mohsen Ahmadi (ahmadim@mail.upb.de) or Prof. Eyke Hüllermeier (eyke@upb.de).

References