Name: Mohd. Wazih Ahmad

Supervisor: Prof. M. N. Doja

Department: Computer Engineering

Title of Thesis: Soft Computing in Intelligent Information Retrieval Systems.

ABSTRACT

Soft computing is an applied area of research to handle the limitations of the traditional computing using new methods from various related fields like Fuzzy Logic, Artificial Neural Network, Machine learning etc. Due to the enormous growth of the internet-based documents, information retrieval is now a basic need of the corporate, institutions, governments and the individuals. In order to handle a large number of candidate documents against a user query, need of the intelligent information retrieval systems exist from the start of this field. To model the uncertainty of deciding the ranks of the individual documents in a search result, various designs for the document representation, indexing systems, and the retrieval functions with application of soft computing techniques were proposed in the past. In this research, limitations of traditional learning to rank methods are addressed and new models for ranking the documents for user queries, based on the experience learned from the past queries are proposed. This research identified the gaps in the traditional rank learning models and proposed two different models to fill those gaps referred as Enumerative Feature Subset Based Ranking System and Compositional Feature Subset Based Ranking System**.** The main limitations of the traditional learning to rank models identified for this research were the absence of the detailed implicit user feedback during learning of the ranking models, the shared feature assumption and the common target ranking assumption. The parameters of the traditional learning to rank model were learned based on the assumption that every user uses all the query-document features to rank the documents, weighted by the parameters of the learned model. But it was found that the different classes of the users do not follow the shared feature assumption while ranking the documents produced by a search engine. Instead, every subset of the given features represents a class of the users. Each class of the users defines its own distribution of the feature subsets with the expected ranks of documents. Also, every class of the user defines its own target ranking as compared to the previous assumption of the single target ranking in the shared

feature model. Each class of the users is modeled as a subset of the features with implicit feedback vector containing eight different types of the feedback parameters. It is assumed that the learning to rank for each class of the users is equivalent to learning a unique pairwise ranking model with specially designed loss function for each feature subset. The loss of a model is inversely proportional to the user feedback its ranking received in the past. All the feature subsets are enumerated as power set of the given feature set and two different probabilistic sampling approaches are proposed. The sample of the feature subsets is used to train the individual models of different cardinalities and finally, individual models are combined together using the ensemble techniques called Bagging and Boosting. In the second approach, each model is learned from a list of enumerated compositions of the given feature set. The feedback obtained by each composition is divided in the proportion of the size of each subset in that composition and a new loss function is proposed for the CFBRS model. The loss function for an ensemble of the EFBRS and the CFBRS models using bagging as a technique of the ensemble of pairwise subset rankers are designed and experiments with the LETOR 4.0 dataset are performed. The proposed rank learning models are learned from the sample of labeled training data using the supervised machine learning technique. The intuition behind learning to rank models is to learn the ranking logic from the past experience and minimize the ranking level loss on the training sample, hoping that the sample queries are representatives of the true population. This research uses empirical risk minimization as a general principle and proved that if the sample is representative of the true population, the set of the optimum parameters of the trained model could be found within the bounds derived on the expected ranking error on the ensemble of the feature subset model. In this case, the proposed models perform well on the full set of the features when compared with the traditional shared feature based ranking model. While on the subset of the features, this model performs better than the traditional model. The proposed model EFBRS has shown comparable performance in terms of the NDCG@10 as compared to the traditional models like RankNet and svmRank for the full set of the features. The loose upper bound on the error of the subset ranking model is proposed for the EFBRS model. The proposed model can be used as an extension to any pairwise and listwise learning to rank model by modifying the respective model for the different cardinalities of the subsets of the features, modifying the corresponding loss function and defining a method to aggregate the individual subset models.