Binary Independence Language Model for Information Retrieval (PI: Dr. Luk Wing Pong Robert; 2009/10)

A new probabilistic retrieval model is proposed for information retrieval. It is called the *binary independence language model* because it is derived from the log-odds ratio of the *binary independence* model, and its probabilities are multiplied and estimated like the statistical *language* model. It is a statistical component model of our probabilistic evaluation model of relevance judgment, so it is a descriptive model that simulates the human relevance decision process. It is consistent with our relevance decision theory and the probabilistic ranking principle that specifies optimal ranking for a variety of retrieval effectiveness measures. Initial results of our model based on terms that are individual words are already better than some existing effective retrieval models (both generative and discriminative ones). Furthering these encouraging results, we propose to enhance the effectiveness of this model by using more sophisticated terms and parameter estimation methods. We focus on establishing good retrieval effectiveness before investigating retrieval efficiency as index design depends on the adopted model for retrieval. Sophisticated terms, including *n*-grams (i.e., contiguous sequences of words), phrases, linked terms (as in dependence language models), class-based *n*-grams, are investigated because they are found to be able to enhance effectiveness by prior research in speech and natural language processing. In that area, discriminative training is found to be helpful too, so it is applied to the parameter estimation of the proposed model. To demonstrate the effectiveness of our model, it will be evaluated by retrospective experiments and in a relevance feedback (RF) environment, because there is current interest in RF, as its results may transfer to implicit feedback from click-through data and as TREC is organizing a RF track with international participation.