Weighting Methods for Information Retrieval Models and Video Retrieval Experiments

論文の内容の要旨

論文題目　Weighting Methods for Information Retrieval Models and Video Retrieval Experiments

（情報検索モデルにおける重み付け法と映像検索データを使用した実証実験）

氏　　名　村田　眞哉

In this dissertation, new information retrieval (IR) models and the video retrieval experiments are addressed. My first two contributions are about the IR model called the Best Match (BM) 25, which is one of the representative and widely-used IR models, and about its application to the video retrieval task called the instance search. The instance search is a challenging task that has been attracting attentions from video retrieval researchers. For this task, given a specific object shown in image queries, a system is developed to rank videos in which the specific object is actually shown. The search results are the list of videos ranked in the decreasing order of their relevance degrees to the specific object. I first experimentally demonstrate that the BM25 with my proposed modification is effective in this task. Such a modification is performed on the discriminative power called the BM25 inverse document frequency (IDF) and I found that enhancing these powers by my methodology significantly improves the instance search accuracy. The new weight is called the exponential IDF (EIDF).

I next show that the EIDF can be theoretically interpreted in the Bayesian framework. In this framework, the setting of the informative prior knowledge leads to enhance the discriminative power and the new weight resembling the EIDF is obtained. Compared with the EIDF, since this formulation is theoretically consistent, the new weight called the Bayesian EIDF (BEIDF) does not retain mathematical problems that the EIDF

suffers from. The retrieval accuracy is also confirmed through the instance search experiments.

The third contribution is regarding the latest IR models called the information-based model (IM) and the divergence from independence (DFI). The term weight for the IM is designed as the extent that the normalized, within-document term frequency diverges from the standard value. The standard value is calculated by the so-called information model and I show that the model based on the generalized Pareto distribution (GPD), which is the main asymptotic distribution in the extreme value statistics (EVS), results in extending the DFI. Together with the novel parameter estimation method for the GPD, the effectiveness of the proposed model is also verified using the instance search dataset.

Since the GPD includes the log-logistic distribution (LLD) as the special case, some existing knowledge on IR models relying on the LLD assumption can be also interpreted from the GPD viewpoint. Since the LLD has been often assumed as the underling distributions when constructing IR models, its extension, that is, GPD, is also expected to become another basic principle in developing new IR models. Exploring this research direction is promising and intriguing.

To summarize, the new IR models are studied and some novel term-weighting methods are derived. Their effectiveness was also experimentally verified using the instance search dataset and I expect that these findings contribute for the further exploration and development of a new family of IR models.