A Maximum Entropy Approach for Modelling Term Dependencies in Probabilistic Information Retrieval

YOU Hyun-Jo Soongsil University
LEE Jung Jin Soongsil University

Introduction

This study presents a probabilistic infromation retrieval model for document ranking task, which is to find an optimal ordered list of documents for a given set of documents and a user query. The proposed approach applies the Maximum Entropy Principle (MEP) to model term dependencies for document ranking according to the Probability Ranking Principle (PRP). The PRP states that documents should ranked by the decreasing probability of relevance to the user request [1], that is, the decreasing conditional probability $\Pr(R = 1|d, q)$, where R is the binary relevance ($R = 1$ if relevant, otherwise $R = 0$), d is a document and q is a query. Documents and queries are represented as binary term incidence vectors. A document is represented as a vector $\boldsymbol{x} = (x_1, \cdots, x_k)$ where $x_j = 1$ if j-th term is present in the document and $x_j = 0$ if the term is absent. A query is represented in same manner. The traditional probabilistic information retrieval models have advocated the unrealistic term independence assumption for practical reason. The so-called Binary Independence Model (BIM) assumes that $\Pr(\boldsymbol{x}|R, \boldsymbol{q}) = \prod_i \Pr(x_i|R, \boldsymbol{q})$ (i.e. Naive Bayes assumption) and plugs in it to the Bayes rule

$$\Pr(R|\boldsymbol{x}, \boldsymbol{q}) = \frac{\Pr(\boldsymbol{x}|R, \boldsymbol{q})P(R|\boldsymbol{q})}{\Pr(\boldsymbol{x}|\boldsymbol{q})} = \frac{\prod_i \Pr(x_i|R, \boldsymbol{q})P(R|\boldsymbol{q})}{\Pr(\boldsymbol{x}|\boldsymbol{q})}$$

and ranks documents by decreasing order of the odds of relevance

$$\frac{\Pr(R = 1|\boldsymbol{x}, \boldsymbol{q})}{\Pr(R = 0|\boldsymbol{x}, \boldsymbol{q})} = \frac{\Pr(R = 1|\boldsymbol{q})}{\Pr(R = 0|\boldsymbol{q})} \prod_{i=1}^{k} \frac{\Pr(x_i|R = 1, \boldsymbol{q})}{\Pr(x_i|R = 0, \boldsymbol{q})}.$$

Lee and Kantor (1991, 1998) proposed a MaxEnt model for information retrieval to estimate $\Pr(\boldsymbol{x}|R, \boldsymbol{q})$, the joint distribution of terms in relevant and nonrelevant documents with known probabilities $\Pr(R|x_i, \boldsymbol{q})$, importance of each term but reported that it was difficult to solve the nonlinear formulation. The present study incorporates an Iterative Proportional Fitting (IPF) algorithm to estimate $\Pr(\boldsymbol{x}|R, \boldsymbol{q})$ and ranks documents by the decreasing order of the posterior probability $\Pr(R = 1|\boldsymbol{x}, \boldsymbol{q})$. Document ranking experiments are presented on data sets from the Microsoft LETOR (LEarning TO Rank) collection.

References

[1] Robertson S.E. (1977). The probability ranking principlein IR. *Journal of Documentation*, Vol. 33, No. 4, 294–304.

[2] Kantor, P.B. and Lee, J.J. (1998). Testing the maximum entropy principle for information retrieval. *Journal of American Society for Information Science*, Vol 49, 6, 557–566.

[3] Lee, J.J. and Kantor, P.B. (1991). A study of probabilistic information retrieval systems in the case of inconsistent expert judgments. *Jounral of American Society for Information Science*, Vol 42, 3, 166–172.