A Latent Variable Model for Query Expansion Using the Hidden Markov Model

Qiang Huang
Knowledge Media Institute
The Open University
Milton Keynes, UK
q.huang@open.ac.uk

Dawei Song
Knowledge Media Institute
The Open University
Milton Keynes, UK
d.song@open.ac.uk

ABSTRACT

We propose a novel probabilistic method based on the Hidden Markov Model (HMM) to learn the structure of a Latent Variable Model (LVM) for query language modeling. In the proposed LVM, the combinations of query terms are viewed as the latent variables and the segmented chunks from the feedback documents are used as the observations given these latent variables. Our extensive experiments show that our method significantly outperforms a number of strong baselines in terms of both effectiveness and robustness.

1. INTRODUCTION

In information retrieval, a prominent query language modeling approach is the Relevance Model (RM) [3]. Practically, variants of RM have shown encouraging performance in ad-hoc search [3, 4, 1] and cross-language retrieval [2], etc.

However, in these existing models, three issues, which in our belief will affect the retrieval effectiveness significantly, have not been addressed. Firstly, there often exist dependencies between query terms (i.e., intra-query term dependencies). Secondly, the distributions $P$ of relevance feedback documents are used as the observations given these latent variables. Our extensive experiments show that our method significantly outperforms a number of strong baselines in terms of both effectiveness and robustness.

2. FRAMEWORK OF THEORY

In our model, the combinations of query terms are seen as the latent variables over the top-ranked documents. The reasons for selecting the query terms as the latent variables are that the word dependence to the query $Q$ is to be estimated in the query language modeling framework, and the estimation of co-occurrence of words with query terms is key to query expansion. Eq. 1 shows an intuitive derivation of theory:

$$P(w|θ_Q) = \sum_{S_j \in S} P(w|S_j)P(S_j) \quad (1)$$

$S_j$ is a latent variable in the set $S (S = \{S_1, \ldots, S_M\})$ and $w$ is the word whose occurrence probability in the expanded query model $θ_Q$ to be estimated. The latent variable $S_j$ is generated from the query $Q$, where each $S_j$ is defined as a combination of query terms. For example, given $Q = \{q_1, q_2\}$, the set of latent variables can be represented as $S = \{\{q_1\}, \{q_2\}, \{q_1, q_2\}\}$. Compared with the traditional methods the use of all the combinations of query terms expands the observing space.

In Eq. 1, the relationship between the word $w$ and the latent variable $S_j$ is derived from the relevance feedback documents. As we have introduced in Section 1, the top-ranked documents in pseudo relevance feedback are not necessarily relevant to the query. Thus, we propose using the segmented chunks to connect $S_j$ and $w$, and the relevance of each chunk to the query will also be considered. Then, a new equation is given as below:

$$P(w|θ_Q) = \sum_{d_i \in d, S_j \in S} P(w|d_i, S_j)P(d_i|S_j)P(S_j) \quad (2)$$

$d$ is the collection of chunks. $P(S_j)$ is the prior distribution of latent variables, $P(d_i|S_j)$ is the probability of an observed chunk $d_i$ given a latent variable $S_j$, and $P(w|d_i, S_j)$ is the probability of a word $w$ in a chunk $d_i$ given a specific latent variable $S_j$.

In order to calculate the three probability parameters in Eq. 2, we design a framework based on the Hidden Markov Model (HMM). The application of the HMM can not only estimate the prior distribution of each $S_j$ but also integrate the dependence between any two latent variables and their underlying observables through a state transition matrix. The details are presented in Figure 1.

Here, we regularize the model estimation with the original query model $P(q_k|Q)$ to alleviate the data sparsity. In this context, the parameters involved in the HMM and the underlying LVM.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

Copyright 200X ACM X-XXXXX-XX-X/XXX...

...$5.00.
paper, the original query model is defined as below:

\[ P(q_k|Q) = \frac{\#q_k \cdot IDF(q_k)}{\sum_{j \in \{1, \ldots |Q|\}} \#q_j \cdot IDF(q_j)} \quad (3) \]

where \( \#q_k \) is the frequency of query term \( q_k \) in \( Q \) and \( IDF(q_k) \) is the inverse document frequency (IDF) of \( q_k \).

To mix the original query model into the newly generated one, two methods are used. HMM-I is an automatic method to integrate the original query model directly into the HMM, in which the original query model as a hidden state \( S_0 \). We therefore can obtain Equation 4.

\[ P_{HMM-I}(w|\theta_Q) = \sum_{d_i \in d} \sum_{S_j \in \{S,S_Q\}} P(w|d_i,S_j)P(d_i|S_j)P(S_j) \quad (4) \]

The second method, HMM-II, uses the linear interpolation to combine the original query model.

\[ P_{HMM-II}(w|\theta_Q) = \lambda P(w|\theta_Q) + (1 - \lambda) P(w|Q) \quad (5) \]

The latter method involves manual adjustment of the interpolation parameter to generate the optimal retrieval performance.

3. EXPERIMENTS AND CONCLUSION

We evaluate our methods by testing TREC topics (Topics151–200, 601–700, and 501–550) on three large collections (AP88–90, ROBUST, and WT10G), respectively. We apply our approach to two scenarios: pseudo-relevance feedback and true relevance feedback. As a comparison, we list the values of MAP obtained by using HMM-II and HMM-I in Table 1.

<table>
<thead>
<tr>
<th>Collection</th>
<th>KL</th>
<th>RM1</th>
<th>RM2</th>
<th>HMM-I</th>
<th>HMM-II</th>
</tr>
</thead>
<tbody>
<tr>
<td>AP88–90</td>
<td>0.2077</td>
<td>0.2603</td>
<td>0.2676</td>
<td>0.2814†</td>
<td>0.2830†</td>
</tr>
<tr>
<td>ROBUST</td>
<td>0.2920</td>
<td>0.3129</td>
<td>0.3143</td>
<td>0.3613†</td>
<td>0.3660†</td>
</tr>
<tr>
<td>WT10G</td>
<td>0.2032</td>
<td>0.2131</td>
<td>0.2134</td>
<td>0.2305†</td>
<td>0.2370†</td>
</tr>
</tbody>
</table>

† The improvement is statistically significant at the level of 0.05 according to the Wilcoxon signed rank test.

Table 1 shows the optimal value when using HMM-II. It is found that using HMM-II can only obtain slightly and insignificantly better performances than HMM-I. However, HMM-II is a somehow heuristic approach compared with HMM-I which integrates original query model neatly in the Hidden Markov Model in a fully automatic way. Therefore, the HMM-I has demonstrated its robustness and effectiveness from both theoretical and practical perspectives.

In this paper, we present a novel method to build a latent variable model using the HMM for query expansion. This paper tries to address some specific issues in those existing models. Our technique aims to incorporate these key issues in a single comprehensive framework and apply the HMM to estimating and optimizing the structure of the LVM. Our experimental results therefore show that our method always obtains significant improvements in comparison with KL, RM1 and RM2.

In terms of future work, it would be very interesting to continue the studies on how to measure the quality of feedback documents and the relevant contents in these documents.

4. REFERENCES