A latent variable model for query expansion using the hidden Markov model

Conference or Workshop Item

How to cite:


For guidance on citations see FAQs.

© 2008 The Authors

Version: Version of Record

Link(s) to article on publisher’s website:
http://dx.doi.org/doi:10.1145/1458082.1458310
http://dl.acm.org/citation.cfm?id=1458310

Copyright and Moral Rights for the articles on this site are retained by the individual authors and/or other copyright owners. For more information on Open Research Online’s data policy on reuse of materials please consult the policies page.

oro.open.ac.uk
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.

Categories and Subject Descriptors
H.3 [Information Storage and Retrieval]: Retrieval Models

General Terms
Algorithm

Keywords
Information retrieval, latent variable model, hidden Markov model

1. INTRODUCTION
In information retrieval, a prominent query language modeling approach is the Relevance Model (RM) [2]. Practically, variants of RM have shown encouraging performance in ad-hoc search [2, 4, 1] and topic detection and tracking [3], 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 \( P(M_j) \), i.e., the importance of different relevance feedback documents, should not be kept uniform. Instead, they should depend on the query terms and \( w \). Thirdly, noise often exists within the feedback documents not any part of a relevant or pseudo-relevant document is necessarily relevant to the query.

The hypothesis of this paper is that tackling all the above identified three issues in a single query language modeling framework will further improve the retrieval effectiveness. In this paper, we first propose segmenting a document into chunks by using an overlapped sliding window. We decompose the query into exhaustive combinations (subsets) of query terms and consider them as latent variables over the chunks. We then propose using a Latent Variable Model (LVM) which connects a chunk \( d \) and a word \( w \) through the latent variables. The dependencies between the latent variables are governed by an ergodic Hidden Markov Model (HMM), where the Viterbi algorithm is applied to optimize parameters involved in the HMM and the underlying LVM.

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:

\[
P(w|\theta_Q) = \sum_{S_j \in S} P(w|S_j)P(S_j) \tag{1}
\]

\( S_j \) 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 \( \theta_{Q} \) 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:

\[
P(w|\theta_Q) = \sum_{d_i \in d \cap S_j \in S} P(w|d_i, S_j)P(d_i|S_j)P(S_j) \tag{2}
\]

\( d \) 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
Figure 1: Outline of framework

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 by combining the query terms. We therefore can obtain Equation 2.

$$P_{HMM-I}(w|\theta_Q) = \sum_{d_i\in D} \sum_{S_J \in [S_S|Q]} P(w|d_i,S_J)P(d_i|S_J)P(S_J)$$ (4)

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

$$P_{HMM-II}(w|\theta_Q) = \alpha P(w|\theta_Q) + (1-\alpha)P(w|Q)$$ (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 (Topics 151–200, 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 1: Comparison of Optimal MAPs using HMM-II and HMM-I

<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>

Pseudo-relevance Feedback

<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>0.3218</td>
<td>0.3427</td>
<td>0.4034</td>
<td>0.4168†</td>
</tr>
<tr>
<td>ROBUST</td>
<td>0.2920</td>
<td>0.4126</td>
<td>0.4432</td>
<td>0.5014</td>
<td>0.5122†</td>
</tr>
<tr>
<td>WT10G</td>
<td>0.2032</td>
<td>0.2571</td>
<td>0.2993</td>
<td>0.3764</td>
<td>0.3859†</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-I 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 on the dependencies between query terms, the different degree of relevance of different chunks in the feedback documents to the query, and noise existing within the feedback documents. Our technique incorporates 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.

4. REFERENCES


