ATLANTIS PRESS
Advances in Computer Science Research (ACRS), volume 54
International Conference on Computer Networks and Communication Technology (CNCT2016)

Exploiting Document Boltzmann Machine in Query Extension

Li-ming HUANG[1], Xiao-zhao ZHAO [2],Yue-xian HOU[3,*] and Ya-ping ZHANG[4]

[1,4] School of Computer Software, Tianjin University, China

[2, 3,*] School of Computer Science & Technology, Tianjin University, China

Keywords: Document Boltzmann Machine, Query Extension , Model Selection, CIF.

Abstract. Most work related to query extension (QE) adopted the assumption that terms in a document are independent, and multinomial distribution is widely used for feedback documents modeling in lots of QE models. We argue that in QE methods, the relevance model (RM) which generates the feedback documents should be modeled with a more suitable distribution, in order to naturally handle the term associations in feedback document. Recently, Document Boltzmann Machine (DBM) was proposed for document modeling in information retrieval, and this model can relax the independence assumption, i.e., can capture the term dependency naturally. It has been shown that DBM can be seen as the generalization of traditional unigram language model and achieves better ad hoc retrieval performance. In this paper, we replace the multinomial distribution in the traditional unigram RM method with DBM, while leaving the main QE framework unchanged to keep the model uncomplicated. Thus, the relevance model is estimated by the DBM trained on feedback documents, called relevance DBM (rDBM). The extended query is generated from the learnt rDBM, and we give the final extended query likelihood according to the parameter values in rDBM. One difficulty in learning rDBM is the problem of data sparseness, which could lead to over fitted rDBMs and harm the retrieval performance. To solve this problem, we adopt Confident Information First (CIF)as model selection principle to reduce the complexity of rDBM, which lead our proposed query extension method more efficient and practical. Experiments on several standard TREC collections show the effectiveness of our QE method with DBM and model selection method.

Introduction

In information retrieval, user express their retrieval needs by a piece of query, e.g., a sentence, or a list of keywords, then the results that conform to users requirements need to be retrieved. However, a lot of important information is lost when expressing information need with the query. Thus there has been a strong interest in exploring the query extension (QE) technique, which can extend the original query to incorporate more implicit information need. Relevance feedback is one of the most popular and elegant QE method. Traditional relevance feedback methods [5][8][9][10] utilize the feedback documents which are relevant or assumed to be relevant, to extend the original query to further improve the performance of retrieval system. To model the feedback documents with a relevance model, these relevance feedback methods utilize the multinomial distribution, which assumes that terms in feedback documents are independent. Obviously, dependence information among terms in feedback documents are missing in such a setting. Although some term association methods can be applied to partially solve this problem, we argue that the relevance model can be better

estimated if the multinomial distribution is replaced with a more suitable distribution, which can naturally abandon the independence assumption.

Recently, the Document Boltzmann Machines (DBMs) [4] are proposed for document modeling, and can achieve better document model and query likelihood theoretically and empirically. DBM is also based on a graph model called all-visible Boltzmann machine, while it can be equipped with principled learning method and inference method. Since the DBM has the ability to model terms and their connection strength naturally, it is a suitable choice for relevance model estimation. We propose to design a framework to apply DBM in query expansion in this paper. We will list some related works in Section 1.1. A brief review of DBM model will be presented in Section 2. The proposed model is described in Section 3, including relevance model estimated by DBM (rDBM) and extended query likelihood calculation based on rDBM. To solve the complexity problem of rDBM, we introduced a parameter selection principle which is very suitable for Boltzmann machines in Section 4. Experiments results are presented in Section 5, and we conclude our work in Section 6.

Review of Document Boltzmann Machines

In this section, we give a brief introduction to the recently proposed document model, Document Boltzmann Machine (DBM). The main procedure of DBM is illustrated in Figure1, which demonstrates the transformation procedure from a document to the corresponding DBM model. At first, a structure of BM called BM template is constructed with each node representingone query term. Based on BM template, each document can be modeled with a BM, called Document BM (DBM). To learn the model for one documentd_i, document segments are sampled from this document and are transformed to vectors which can be used as input of the BM template. Specifically, in the vector of one segment, the dimensions where the corresponding terms appear areassigned to 1, and 0 otherwise. More detailed document segments sampling method can be find in [4].

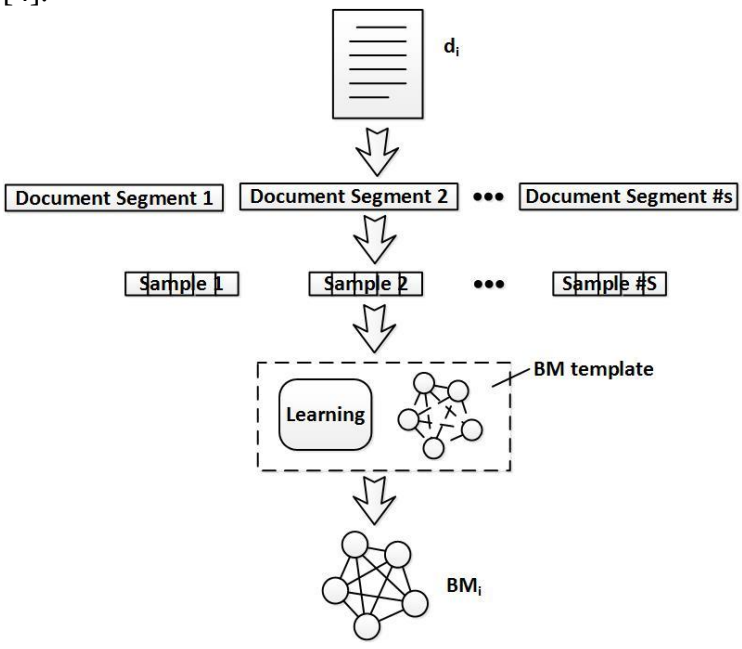

Figure 1. Document Boltzmann Machine

With the set of sampled vectors X, the corresponding document can be represented as a DBM via maximum likelihood (ML) learning method. The learning objective can be written as:

$$\log p(\mathrm{x}; W) = \quad -E(\mathrm{x}; W) - \log[Z(W)] \tag{1}$$

where W is the parameter set of Boltzmann machine(W contains 1-order parameters w^1 and 2-order parameters w^2), $E(W)$ is the energy function:

$$\mathrm{E}(\mathrm{x}; \mathrm{W}) = E(\mathrm{x};\ w^1, w^2) = \ -\ \sum_i w_i^1 x_i - \ \sum_{i<j} w_{ij}^2 x_i x_j \tag{2}$$

and $\mathrm{Z}(\mathrm{W}) = \ \sum_x \exp[-E(x; W)]$ is the partition function.

To rank documents, query likelihood from each DBM model is calculated via

$$\log p\left(\mathrm{x}_Q \middle| BM\right) = \log p\left(\mathrm{x}_Q; W\right) \tag{3}$$

Besides, related smoothing method is also proposed and empirical test of this DBM retrieval system reveals the advantage of this new distribution assumption compared to traditional unigram language model.

Exploiting DBM in Query Extension

In this section, we describe our retrieval framework of query expansion using DBM model. Specifically, we proposed a new method to model relevance information in feedback documents in Section 3.1. Then the whole procedure of our query expansion will be described in Section 3.2.

Modeling the Relevance Model by BM

Relevance model is a critical concept in information retrieval, and lots of methods have been proposed for it. In [5], a well-known RM method is proposed based on language model [1], which model the relevance information with a multinomial distribution based on feedback documents. Since DBM has the same functionality as LM to model documents, it can also be used to model the relevance information. We call the DBM estimated relevance model as Relevance DBM (rDBM). Here we give the specific steps tolearning arDBM fromfeedback documents.The procedure of the rDBM modeling is shown in Figure 2. Please note that the sub-procedure in the dashed rectangle will be illustrated in Section 3.1, and it is optimal for our model.

When we use the DBM to model the relevance model based on feedback document, the first problem is how to decide what does each node in DBM stands for. In our work, we care more of the most important terms in feedback document and the original query. Specifically, we select k terms with top TF-IDF from the (pseudo) feedback documents, and these terms and the original query terms are represented by the nodes of the BM template. Then all the BMs in our framework are learned based on this BM structure.

Before training our BM based relevance model, we need to prepare the document segment samples from the feedback documents. With all the feedback documents, we use a sliding window to get overlapped feedback documents segments with a fixed window size σ and step length of the sliding α.These segments are represented as binary vectors with the method described in Section 2. This sampling method is the same as the one used in [4].These segments sampled from the feedback documents contains both frequency and dependency information of terms in relevance model, so the BM model learned based on these samples can provide a better relevance model estimation, while in traditional relevance model, multinomial distribution lack the ability for dependency modeling.

With both the BM template and segment samples prepared, we are ready to train the BM model. With the simple likelihood maximization training methods, we can obtain a BM model which represent the relevance information from the sampled segments. We call this DBM based relevance model as rDBM, which play a critical role in our query extension method.

The section headings are in boldface capital and lowercase letters. Second level headings are typed as part of the succeeding paragraph (like the subsection heading of this paragraph).

The second and following paragraph under the first and second level titles (the Introduction part is excluded) need to indented for one character. Please do not forget.

Query Expansion with Relevance BM

Once we get the DBM relevance modeling from feedback documents, namely the rDBM, we can conduct the query extension methods. Instead of giving the completed weighted query, we directly show the final query likelihood, since such a transforming is simple and the final formulation is more straightforward for implementation and explanation.

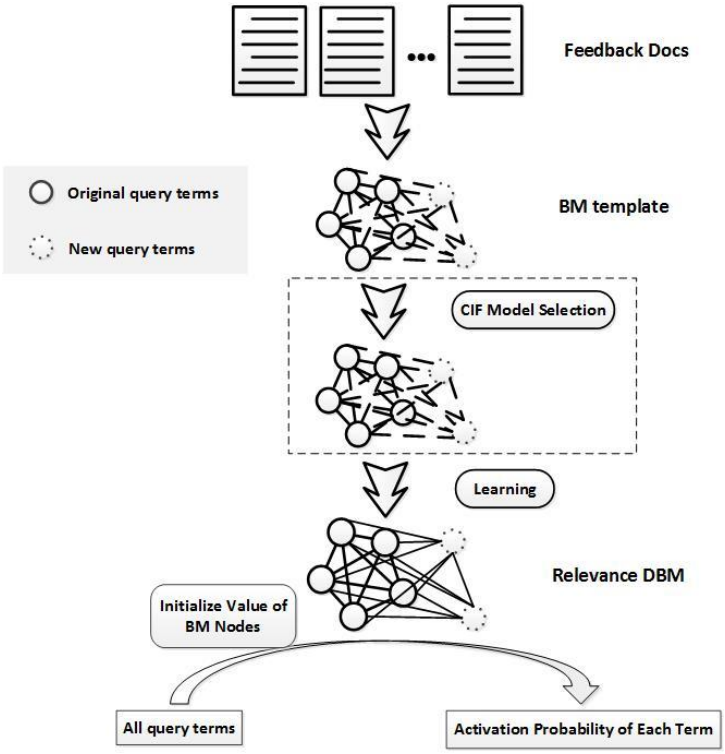

Figure 2. Query Expansion with Relevance BM

To exploit our learned rDBM for document ranking, we rewrite the query likelihood of each document as:

$$\mathrm{p(Q|d)} = \lambda_1 \sum_{q_i \in Q_0} \log p(q_i|d) + \lambda_2 \sum_{e_j \in E} \alpha^1_{e_j} \log p(e_j|d) + \lambda_3 \sum_{v_i \in Q_0, v_j \in E \cup Q_0} \alpha^2_{w_i w_j} \log p(v_i v_j|d) \quad (4)$$

where $\alpha^1_{e_j}$ and $\alpha^2_{v_i v_j}$ are corresponding weight which will be defined in the following. In this formulation, Qare all query terms,Q_0 are the original query terms and E are the new query terms. Namely, $\mathrm{Q} = \{Q_0, \mathrm{E}\}$. The three terms in the above equation are weighted by λ_1, λ_2 andλ_3 just like in MRF model [2] and their sum is 1.

Now we describe how to obtain the term weight $\alpha_{e_j}^1$and $\alpha_{v_i v_j}^2$. Firstly, $\alpha_{e_j}^1$ is inferenced by activation equation of rDBM. Specifically, the original query is used to initialize the value of rDBM nodes, namely, we set the value of nodes for original query terms to 1 and other nodes to 0. Starting from this node value, all the nodes can be activated with a probability. The activation probability of the node e_j (standing for an expansion query term in rDBM) can be written as:

$$\alpha_{e_j}^1 = P\left(V_{e_j} = 1\right) = \frac{1}{1 + exp[-E(e_j)]} \tag{5}$$

where $E(e_j)$ is the energy function mentioned before.

Then for the two-term combination weights $\alpha_{v_i v_j}^2$, we utilize the 2-order parameter in the Boltzmann distribution formulation as their weight. This strategy is simple and meaningful since the parameter in BM can beused to measure the association strength among nodes/terms. Thus, we can give strong term associations more influence. Formally, we define a weight when calculating the query likelihood for node $v_i$andnodev_j:

$$\alpha_{v_i v_j}^2 = w_{ij} \tag{6}$$

With this query likelihood, we combine the original query terms and extended query terms with proper and automatically learned weight. This extended query not only considers the single words, but also the associations between query terms. We can get the weight $\alpha_{v_i v_j}^2$ between query term w_i and w_j from the Relevance DBM. In other words, the extended query includes three parts: one part are the original query terms which have the same weights; another part are the new extended terms which have the different weights; the last part are any combination of two terms.

The rDBM model for feedback document naturally models the dependencies among terms. These associations influence the final query via give more weights to the terms and term associationswhich are more dependent on the original query terms.

Discussion of rDBM and LCE Model

There are another graphical model designed for query extension, the Latent Concept Expansion (LCE) [3] model. LCE bring the basic idea of MRF to query expansion to model term dependencies as well as make use of arbitrary text features. However, LCE needs to manually set all kinds of feature weighting, which makes it very parameter sensitive and increases the model implementation burden. So there is great meaning to propose rDBM model that can model the query terms and associations among query terms via automatically learning procedure.Besides, the final query likelihood of rDBM considers the meaning of parameters in BM model, which is also more formal than LCE, without manually parameter value assignment or parameter tuning with validation set.

Improve the Efficiency of rDBM in Query Expansion

As shown in Figure 2, the query expansion method adds more terms into the expanded query, and the structure of the BM template has been became complex accordingly, i.e., more parameters are involved. Thus, one difficulty in learning rDBM is the problem of data sparseness, which could lead to overfitted model and harm the retrieval performance. This problem will also urge us to design more efficient learning method, which may be critical to the success of the application of DBM in practice. To solve

these problems, we propose to reduce the complexity of rDBM by an efficient model selection algorithm for Boltzmann machines.

Recently, A confident information First (CIF) Principle [6] is proposed for parameter reduction criterion. The proposed CIF is fundamentally different from the traditional feature reduction method. CIF can be used to guide the model selection with the theoretical foundation in parameter space, and is a more general principle.Empirical result also demonstrated the advantage of CIF compared to traditional model selection method such as AkaikeInformation Criterion (AIC) [7].Thus, we choose this principle to guide our model simplification of rDBM.

Model Selection for rDBM

The CIF principle maximally conserve confident parameters and remove less confident parameters. The confidence of every parameter is evaluated by its contribute to the expected Fisher information distance within the geometric manifold over the neighborhood of the underlying distribution. It is worth emphasizing that the CIF principle can reduce the time complexity of the model selection procedure for BM. Moreover, there is an elegant algorithm proposed in [6] (see Algorithm 1 in [6]). This algorithm give a model selection procedure for BM without manually assigned parameter, which appeals since we hope to avoid introducing extra parameters into our rDBM framework.

In addition to speeding up the learning procedure, the procedure of model selection can be regarded as a regularization for the relevance DBM, which is important for DBM learning, e.g., smoothing method in [4]. Without this model selection as regularization, there are a large amount of parameters in the relevance DBM, while the segments data are very sparse. This may cause model overfitting since the number of parameter in relevance DBM is too large for training samples from feedbackdocuments. Specially, as shown in Figure 2, we will get the new BM template, when simplifying the BM template by CIF (denoted by the dashed rectangle). CIF reduce the number of parameters to acceptable degree and can preserve the most information of the left parameters. This property of CIF can regularize the rDBM to get rid of parameters with low estimation confidence. Since CIF can help rDBM to have a higher average estimation confidence of parameters, we believe that CIF can help rDBM to achieve better performance as well as to improve the efficiency.

Experiments

In this section, we implement and test the methods proposed in this paper. We empiricallycompare our methods with classic query expansion method to test its effectiveness. The performance of model selection algorithm designed for BM is also evaluated.

Datasets and Settings

We used four TREC datasets in our experiments: AP8889 (query 151-200, 164597 documents), WSJ8792 (query 151-200, 173252 documents), ROBUST2004 (query 601-700, 528155 documents), WT10G (query 501-550, 1692096 documents). We build the index of these collections using Indri 5.7, with all words stemmed by Porter stemmer and stopwords removed. In all the experiments, the length of extended query is set to 100, and we utilize the top 50 documents as feedback documents. We record the

MAP (Mean Average Precision), P@5 and P@10, to compare performance of different methods.

DBM based Query Expansion

At first, we test our query expansion method rDBM, without CIF model selection. We choose the classic RM method as the baseline. These two methods are both query expansion and share the same basic framework, except they estimating the relevance model with different kinds of distribution.

Table 1.Evaluation of rDBM (performance changes are based on RM)

	AP8889			WSJ8792		
Metric	MAP	P@5	P@10	MAP	P@5	P@10
LM	0.2016	0.6279	0.5280	0.3169	0.7611	0.7064
DBM	0.2074	0.6424	0.5385	0.3200	0.8144	0.7531
RM	0.2732	0.6648	0.6050	0.3740	0.8425	0.7734
rDBM	0.2776 (+1.62%)	0.6891 (+3.66%)	0.6179 (+2.13%)	0.3781 (+1.11%)	0.8672 (+2.93%)	0.7875 (+1.71%)
rDBM(CIF)	0.2778 (+1.68%)	0.6913 (+3.99%)	0.6181 (+2.17%)	0.3782 (+1.13%)	0.8722 (+3.53%)	0.7886 (+1.84%)
	ROBUST2004			WT10G		
Metric	MAP	P@5	P@10	MAP	P@5	P@10
LM	0.2802	0.8393	0.7433	0.1831	0.3637	0.3484
DBM	0.2839	0.8589	0..7585	0.1852	0.3754	0.3582
RM	0.3388	0.8127	0.7601	0.1965	0.3805	0.3676
rDBM	0.3422 (+1.01%)	0.8321 (+2.38%)	0.7735 (+1.76%)	0.1990 (+1.27%)	0.3839 (+0.89%)	0.3715 (+1.06%)
rDBM(CIF)	0.3425 (+1.09%)	0.8402 (+3.38%)	0.7751 (+1.98%)	0.1977 (+0.61%)	0.3816 (+0.39%)	0.3692 (+0.44)

Experiments results are shown in Table 1. At first, we can observe that DBM perform better than LM, which is consistent with the results reported [4]. RM improve the performance significantly compared to both LM and DBM, since feedback documents provide lots of relevance information and query expansion is a very useful method for information retrieval. Table 1 also demonstrates that our query expansion method with rDBM is more effective than RM to improve the retrieval performance. This result attributes to the advantage of the Boltzmann distribution used in query expansion, and confirms that the combination of DBM and query expansion is practical.

Model Selection for DBM in Query Expansion

In this section, we test the model selection method and give the comparison result between rDBM based query expansion with and without model selection.
Performance of rDBM based query expansion with the model selection method (CIF) are recorded in Table 1. According to these results, we find that CIF can help to improve the retrieval performance of rDBM based query expansion, and this demonstrates the ability of CIF to solve overfitting problem and regularize theoriginal rDBM. In addition to the better performance, the rDBM method with CIF is more efficient since the number of parameters in rDBM template are reduced based on CIF principle.Consistent to the our discussion in Section 4, the model selection indeed can reduce the learning time of rDBM and improve the performance as well.

Conclusion and Future Work

At the first time, we exploit DBM model in query expansion. We model the relevance feedback document with a DBM, named rDBM, as the relevance model. With this

automatically learned relevance model, we can extract parameter values to help to expand the original query with proper weight, and consequently the proper likelihood of extended query. This query expansion method based on rDBM is intuitive and has proper theoretical explanation. In addition, to solve both the model overfitting and efficiency problems, Confidence Information First(CIF) principle is utilized to further modify the rDBM template, which can accelerate the learning procedure of rDBM as well as improve performance with better estimated parameters in rDBM. Experiments on four standard TREC datasets demonstrate the effectiveness of rDBM based query expansion and its CIF regularization.

For future work, we would continue to theoretically justify the parameter usage in rDBM based query expansion. Besides, before we apply rDBM to other application scenarios, we would test this method on more datasets and compare it with more competitive baseline.

References

[1] Ponte J M, Croft W B. A Language Modeling Approach to Information Retrieval[J]. Research & Development in Information Retrieval, 1998:275–281.

[2] Metzler D, Croft W B. A Markov random field model for term dependencies[C]// SIGIR 2005: Proceedings of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, August. 2005:472-479.

[3] M etzler D, Croft W B. Latent concept expansion using markov random fields.[C]// SIGIR 2007: Proceedings of the, International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, the Netherlands, July. 2007:311-318.

[4] Y Q, Zhang P, Hou Y, et al. Document Boltzmann Machines for Information Retrieval[M]. Advances in Information Retrieval. Springer International Publishing, 2015:666-671.

[5] Lavrenko V, Croft W B. Relevance based language models[C]. International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2010:120-127.

[6] Zhao X, Hou Y, Song D, et al. A Confident Information First Principle for Parametric Reduction and Model Selection of Boltzmann Machines[J]. Japanese Journal of Physiology, 2015, 9(1):282-303.

[7] Akaike H. A New Look at the Statistical Model Identification[J]. Automatic Control IEEE Transactions on, 1974, 19(6):716-723.

[8] R occhio J. Relevance Feedback in Information Retrieval[M]// The SMART Retrieval System: Experiments in Automatic Document Processing. 1971:313-323.

[9] Z hai C, Lafferty J. Model-based Feedback in the Language Modeling Approach to Information Retrieval[C]// Tenth International Conference on Information & Knowledge Management. 2001:403-410.

[10] Xu J, Croft W B. Improving the effectiveness of information retrieval with local context analysis[J]. Acm Transactions on Information Systems, 2000, 18(1):79-112.