A Discriminative Approach to Topic-based Citation Recommendation *

Jie Tang and Jing Zhang

Department of Computer Science and Technology,
Tsinghua University, Beijing, 100084. China
jietang@tsinghua.edu.cn, zhangjing@keg.cs.tsinghua.edu.cn

Abstract. In this paper, we present a study of a novel problem, i.e. topic-based citation recommendation, which involves recommending papers to be referred to. Traditionally, this problem is usually treated as an engineering issue and dealt with using heuristics. This paper gives a formalization of topic-based citation recommendation and proposes a discriminative approach to this problem. Specifically, it proposes a two-layer Restricted Boltzmann Machine model, called RBM-CS, which can discover topic distributions of paper content and citation relationship simultaneously. Experimental results demonstrate that RBM-CS can significantly outperform baseline methods for citation recommendation.

1 Introduction

Citation recommendation is concerned with recommending papers that should be referred to. When starting a work in a new research topic or brainstorming for novel ideas, a researcher usually wants to have a quick understanding of the exiting literatures in this field, including which papers are the most relevant papers and what sub-topics are presented in these papers. Two common ways to find reference papers are: (1) search documents on search engines such as Google and (2) trace the cited references by starting with a small number of initial papers (seed-papers). Unfortunately, for (1) it would be difficult to find a comprehensive keyword list to cover all papers, especially for beginning researchers. It is very possible to miss important developments in areas outside a researcher's specialty. For (2), an average paper may cite more than twenty papers. It would be quite time consuming to analyze each of the cited reference to see whether it is useful or not, especially with the increase of the tracing depth. Additionally, even a well organized paper may miss some important "related work", due to space limitation or other reasons.

Previously, papers recommendation has been studied, for example, by exploring collaborative filtering [7]. Our problem is relevant, but different, from this kind of work. Firstly, in citation recommendation, the user is interested in not only a list of recommended papers, but also the sub-topics presented in these papers. Secondly, conventional methods can only recommend papers; but cannot suggest the citation position (i.e., which sentences should refer to the citation).

* The work is supported by the National Natural Science Foundation of China (60703059), Chinese National Key Foundation Research and Development Plan (2007CB310803), and Chinese Young Faculty Research Funding (20070003093).

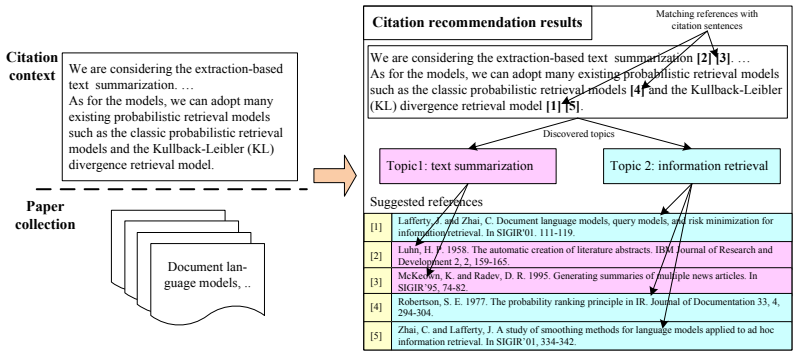

Fig. 1: Example of citation recommendation.

In this paper, we formalize citation recommendation as that of topic discovery, topic-based recommendation, and matching citation sentences with the recommended papers. We propose a unified and discriminative approach to citation recommendation. This approach can automatically discover topical aspects of each paper and recommend papers based on the discovered topic distribution. Experimental results show that the proposed approach significantly outperforms the baseline methods.

2 Problem Formulation

We define notations used throughout this paper. Assuming that a paper d contains a vector $\mathbf{w}_d$ of N_d words, in which each word w_{di} is chosen from a vocabulary of size V; and a list $\mathbf{l}_d$ of L_d references. Then a collection of D papers can be represented as $\mathbf{D} = \{(\mathbf{w}_1, \mathbf{l}_1), \cdots, (\mathbf{w}_D, \mathbf{l}_D)\}$. We only consider references that appear in the paper collection $\mathbf{D}$. Thus the size L of the vocabulary of references is D. Further, we consider that each paper is associated with a distribution of T topics, so is the citation.

Definition 1. ***(Citation Context and Citation Sentence)*** *Citation context is defined by the context words occurring in, for instance, the user written proposal. For an example, the words "... We use Cosine computation [x] evaluate the similarity ..." would be a citation context. One reference paper is expected to be cited at the position "[x]". We use c to denote a citation context. Each sentence in the citation context is called citation sentence. The position "[x]" to cite the reference paper is called citation position.*

Figure 1 shows an example of citation recommendation. The left part of Figure 1 includes a citation context provided by the user and a paper collection. The right part shows the recommended result that we expect a citation recommendation algorithm outputs. For instance, two topics, i.e., "text summarization" and "information retrieval", have been extracted from the citation context. For the first topic "text summarization", two papers have been recommended and for the second topic "information retrieval", three papers have been recommended. Further, the recommended papers are matched with the citation sentences and the corresponding citation positions have been identified.

We see that the recommended papers are topic dependent. By nature, the problem of citation recommendation can be formalized as topic discovery, reference papers recommendation, and matching of the recommended papers with the citation sentences.

3 Our Approach

At a high level, our approach primarily consists of three steps:

1. We propose a two-layer Restricted Boltzmann Machine (RBM) model, referred to as RBM-CS. Given a collection of papers with citation relationship, the model learns a mixture of topic distribution over paper contents and citation relationships.
2. We present a method to rank papers for a given citation context, based on the learned topic model. We take the top ranked papers as the recommended papers.
3. We describe a method to find the correspondence between the recommended papers and the citation sentences.

3.1 The RBM-CS Model

Restricted Boltzmann Machines (RBMs) [8] are undirected graphical models that use a layer of hidden variables to model a (topic) distribution over visible variables. In this work, we propose a two-layer RBM model, called RBM-CS, to jointly model papers and citations. Graphical representation of the RBM-CS model is shown in Figure 2. We see that in RBM-CS, the hidden layer **h** is associated with two visible layers: words **w** and citation relationships **l**, respectively coupling with an interaction matrix **M** and **U**. The basic idea in RBM-CS is to capture the topic distribution of papers with a hidden topic layer, which is conditioned on both words and citation relationships. Words and citation relationship are considered to be generated from the hidden topics independently.

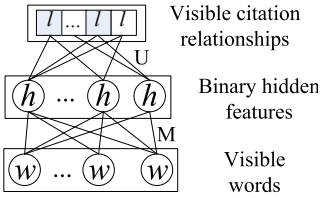

Fig. 2: Graphical representation of the RBM-CS model.

To train a graphical model, we can consider maximization of the generative log-likelihood $\log p(\mathbf{w}, \mathbf{l})$. However, we are dealing with a predictive problem, our interests ultimately only lie in correct prediction $p(\mathbf{l}|\mathbf{w})$, not necessarily to have a good $p(\mathbf{w})$. Therefore, we define a discriminative objective function by a conditional log-likelihood:

$$L = \sum_{d}^{D} \log p(\mathbf{l}_d|\mathbf{w}_d) = \sum_{d}^{D} \log \left(\prod_{j=1}^{L} p(l_j|\mathbf{w}_d) \right) \tag{1}$$

The probability $p(l_j|\mathbf{w}_d)$ can be defined as:

$$p(l_j|\mathbf{w}) = \sigma\left(\sum_{k=1}^{T} U_{jk} f(h_k) + e_j\right), f(h_k) = \sigma\left(\sum_{i=1}^{V} M_{ij} f(w_i) + \sum_{j} U_{kj} f(l_j) + a_k\right) \quad (2)$$

where $\sigma(.)$ is a sigmoid function, defined as $\sigma(x) = 1/(1 + exp(-x))$; e are bias terms for citation relationships; $f(h_k)$ is the feature function for hidden variable h_k; $f(l_j)$ and $f(w_i)$ are feature functions for citation relationship l_j and word w_i, respectively; a are bias terms for hidden variables. For simplicity, we define $f(w_i)$ as the count of word w_i in document d. We define binary value for the feature function of citation relationship l. For example, for document d, $f(l_j) = 1$ denotes that the document d has a citation relationship with another paper d_j.

Now, the task is to learn the model parameters $\Theta = (\mathbf{M}, \mathbf{U}, \mathbf{a}, \mathbf{b}, \mathbf{e})$ given a training set $\mathbf{D}$. Maximum-likelihood (ML) learning of the parameters can be done by gradient ascent with respect to the model parameters ($\mathbf{b}$ are bias terms for words). The exact gradient, for any parameter $\theta \in \Theta$ can be written as follows:

$$\frac{\partial \log p(\mathbf{l}|\mathbf{w})}{\partial \theta} = E_{P_0}[\mathbf{l}|\mathbf{w}] - E_{P_M}[\mathbf{l}|\mathbf{w}] \quad (3)$$

where $E_{P_0}[.]$ denotes an expectation with respect to the data distribution and E_{P_M} is an expectation with respect to the distribution defined by the model. Computation of the expectation E_{P_M} is intractable. In practice, we use a stochastic approximation of this gradient, called the contrastive divergence gradient [4]. The algorithm cycles through the training data and updates the model parameters according to Algorithm 1, where the probabilities $p(h_k|\mathbf{w}, \mathbf{l})$, $p(w_i|\mathbf{h})$ and $p(l_j|\mathbf{h})$ are defined as:

$$p(h_k|\mathbf{w}, \mathbf{l}) = \sigma(\sum_{i=1}^{V} M_{ik} f(w_i) + \sum_{j=1}^{L} U_{jk} f(l_j) + a_k) \quad (4)$$

$$p(w_i|\mathbf{h}) = \sigma(\sum_{k=1}^{T} M_{ik} f(h_k) + b_i) \quad (5)$$

$$p(l_j|\mathbf{h}) = \sigma(\sum_{k=1}^{T} U_{jk} f(h_k) + e_j) \quad (6)$$

where b are bias terms for words; $f(l_j)$ is the feature function for citation relationship.

Algorithm 1. Parameter learning via contrastive divergence

Input: training data $\mathbf{D} = \{(\mathbf{w}_d, \mathbf{l}_d)\}$, topic number T, and learning rate λ

1. repeat
   (a) for each document d:
      i. sampling each topic h_k according to (4);
      ii. sampling each word w_i according to (5);
      iii. sampling each citation relationship l_j according to (6);
   (b) end for
   (c) update each model parameter $\theta \in \Theta$ by

$$\theta = \theta + \lambda(\frac{\partial \log p(\mathbf{l}|\mathbf{w})}{\partial \theta})$$

2. until all model parameters Θ converge

3.2 Ranking and recommendation

The objective of citation recommendation is to rank the recommended papers for a given citation context. Specifically, we apply the same modeling procedure to the citation context. Hence, we can obtain a topic representation $\{\mathbf{h}_c\}$ of the citation context c. Based on the topic representation and the modeling results, we can calculate the probability of each paper being the reference paper for the citation context according to Equation (6). Finally, the papers are ranked in terms of the probabilities and the top K ranked papers are returned as the recommended papers. It is hard to specify an accurate value of K for each citation context. A simple way is to set it as the average number of citations in a paper (i.e., 11 in our data set).

3.3 Matching Recommended Papers with Citation Sentences

The purpose of matching the recommended papers with citation sentences is to align the recommended papers with sentences in the citation context. This can be done by using each recommended paper as a keyword query to retrieve relevant citation sentences. In general, we may use any retrieval method. In this paper, we used KL-divergence to measure the relevance between the recommended paper and the citation sentence:

$$KL(d, s_{ci}) = \sum_{k=1}^{T} p(h_k|d) \log \frac{p(h_k|d)}{p(h_k|s_{ci})} \tag{7}$$

where d is a recommended paper and s_{ci} is the ith sentence in the citation context c; the probabilities $p(h_k|d)$ and $p(h_k|s_{ci})$, which can be obtained by (4).

4 Experiments

4.1 Experimental Setting

Data Set We conducted experiments on two data sets, NIPS [1] and Citeseer [2]. The NIPS data set consists of 12 volumes of NIPS papers (1,605 papers and 10,472 citation relationships). Each paper contains full text and its citations. We removed some citations with incomplete information, e.g., consisting of only authors and publication venue, but no title. We also removed citations that do not appear in the data set. The Citeseer data set consists of $3,335$ papers (with 32,558 citation relationships) downloaded from the Citeseer web site. As well, we removed citations that do not appear in the data set.

Each paper was preprocessed by (a) removing stopwords and numbers; (b) removing words appearing less than three times in the corpus; and (c) downcasing the obtained words. Finally, we obtained $V = 26,723$ unique words and a total of $350,361$ words in NIPS and $V = 44,548$ unique words and $634,875$ words in Citeseer.

[1] http://www.cs.toronto.edu/~roweis/data.html
[2] http://citeseer.ist.psu.edu/oai.html

Table 1: Two topics discovered by RBM-CS from the NIPS data.

"Topic 12: Markov Model"			
Words		Citation	
hmm	0.091	links between Markov models and multilayer perceptrons	0.0347
state	0.063	a tutorial on hidden Markov models and selected applications in speech recognition	0.0221
markov	0.058	connectionist speech recognition a hybrid approach	0.0169
probability	0.057	global optimization of a neural network hidden Markov model hybrid	0.0169
field	0.018	neural network classifiers estimate Bayesian a posteriori probabilities	0.0169
"Topic 97: Support Vector Machines"			
Words		Citation	
kernel	0.083	the nature of statistical learning	0.036363
margin	0.079	a training algorithm for optimal margin classifiers	0.026984
support	0.075	a tutorial on support vector machines for pattern recognition	0.026763
svm	0.075	statistical learning theory	0.020220
machine	0.069	support vector networks	0.015117

Evaluation Measure and Baseline Methods We used P@1, P@3, P@5, P@10, Rpec, MAP, Bpref, and MRR as the evaluation measures. For the details of the measures, please refer to [1] [2]. We conducted the evaluation on both paper-level (without considering the citation position) and sentence-level (considering the citation position).

We defined two baseline methods. One is based on language model (LM). Given a citation context c, we computed the score of each paper d by $p(c|d) = \prod_{w \in c} p(w|d)$, where $p(w|d)$ is the maximum likelihood of word w in document d. We ranked papers according to this score and recommended the top K ranked papers.

The other baseline is based on RBM, which learns a generative model for papers and the citation context. Then we use KL-divergence to calculate a score for each paper (by a similar equation to Equation (7)). For both RBM and RBM-CS, we set the number of topic as $T = 200$ and the number of recommended references as the average number of the data set, i.e. $K = 7$ for NIPS and $K = 11$ for Citeseer. The weights were updated using a learning rate of 0.01/batch-size, momentum of 0.9, and a weight decay of 0.001.

4.2 Experimental Results

Estimated Topics Table 1 shows two example topics discovered by RBM-CS from the NIPS data. We can see that our model can capture the topic distribution very well.

Performance of Citation recommendation Table 2 shows the result of citation recommendation. We see that our proposed model clearly outperforms the two baseline models. The language model suffers from the fact that it is based on only keyword matching. The RBM uses a hidden topic layer to alleviate the problem. However, it is aimed at optimize $p(\mathbf{w})$, which might be inappropriate for citation recommendation. In addition, RBM cannot capture the dependencies between paper contents and citation relationships. Our proposed RBM-CS can be advantageous to optimize $p(l|w)$ directly and to model the dependencies between paper contents and citation relationships.

We can also see from Table 2 that the recommendation performance is much better on the Citeseer data than that on the NIPS data. This means that on the sparse data, the recommendation tasks would be more difficult. How to improve the recommendation performance on the sparse data is also one of our ongoing work.

Table 2: Performance of citation recommendation on the two data sets.

Data	Method	P@1	P@2	P@3	P@5	P@10	Rprec	MAP	Bpref	MRR
NIPS	LM	0.0195	0.0164	0.0132	0.0125	0.0148	0.0161	0.0445	0.0108	0.0132
	RBM	0.0289	0.0313	0.0263	0.0224	0.0164	0.0245	0.0652	0.0176	0.0162
	RBM-CS	0.2402	0.2628	0.2349	0.1792	0.1170	0.1676	0.3499	0.1626	0.1082
Citeseer	LM	0.0496	0.0492	0.0454	0.0439	0.0274	0.0259	0.1103	0.0311	0.0243
	RBM	0.1684	0.1884	0.1780	0.1519	0.0776	0.1510	0.2804	0.1189	0.0639
	RBM-CS	0.3337	0.3791	0.3501	0.2800	0.1768	0.2375	0.4237	0.2501	0.1564

Table 3: Performance of sentence-level citation recommendation on the NIPS data set.

Model	P@1	P@2	P@3	P@5	P@10	Rpec	MAP	Bpref	MRR
LM	0.0783	0.0642	0.0582	0.0629	0.00503	0.0607	0.1178	0.0483	0.0502
RBM	0.1081	0.1061	0.1061	0.1000	0.0727	0.0914	0.2089	0.0761	0.0851
RBM-CS	0.2005	0.2136	0.2010	0.1788	0.1561	0.1782	0.2854	0.1565	0.1657

Table 3 shows the performance of citation recommendation by RBM and RBM-CS in terms of sentence-level evaluation. (As the Citeseer data contains a lot OCR errors and it is difficult to accurately extract the citation position, we conducted sentence-level evaluation on the NIPS data only.) We can again see that our proposed model significantly outperforms the method of using LM and that of using RBM.

5 Related Work

We review scientific literatures about citation analysis and related topic models. Citation analysis usually employs a graphical model to represent papers and their relationships, for example Science Citation Index [3]. This index links authors and their corresponding papers. Bibliographical Coupling (BC) [6] and co-citation analysis are proposed for citation analysis, for example to measure the quality of an academic paper [3].

Recommending citations for scientific papers is a task which has not been studied exhaustively before. Strohman et al. [9] investigated this task using a graphical framework. Each paper is represented by a node and the citation relationship is represented as the link between nodes. A new paper is a node without in and out links. Citation recommendation is then cast as link prediction. McNee et al. [7] employed collaborative filtering in citation network to recommend citations to papers. Both of them use the graphical framework. We look at citation recommendation from a different perspective. We take advantages of the dependencies between paper contents and citation relationships by using a hidden topic layer to joint model them.

Restricted Boltzmann Machines (RBMs) [8] are generative models based on latent (usually binary) variables to model an input distribution, and have been applied in a large variety of problems in the past few years. Many extensions of the RBM model have been proposed, for example dual wing RBM [12], modeling various types of input distribution [5] [11]. In this paper, we propose a two-layer Restricted Boltzmann Ma-

chine model, called RBM-CS, which can jointly model topic distribution of papers and citation relationships.

6 Conclusion

In this paper, we formally define the problems of topic-based citation recommendation and propose a discriminative approach to this problem. Specifically, we proposes a two-layer Restricted Boltzmann Machine model, called RBM-CS, to model paper contents and citation relationships simultaneously. Experimental results show that the proposed RBM-CS can significantly improve the recommendation performance.

There are many potential future directions of this work. It would be interesting to include other information for citation recommendation, such as conference and author information. We are going to integrate the citation recommendation as a new feature into our academic search system ArnetMiner [10] (http://arnetminer.org).

References

1. C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'04)*, pages 25–32, 2004.
2. N. Craswell, A. P. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *TREC 2005 Conference Notebook*, pages 199–205, 2005.
3. E. Garfield. Citation analysis as a tool in journal evaluation. *Science*, 178(4060):471–479, 1972.
4. G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14:1771–1800, 2002.
5. G. E. Hinton. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.
6. M. M. Kessler. Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25, 1963.
7. S. M. McNee, I. Albert, D. Cosley, P. Gopalkrishnan, S. K. Lam, A. M. Rashid, J. A. Konstan, and J. Riedl. On the recommending of citations for research papers. In *CSCW'02*, pages 116–125, 2002.
8. P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. pages 194–281, 1986.
9. T. Strohman, W. B. Croft, and D. Jensen. Recommending citations for academic papers. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07)*, pages 705–706, 2007.
10. J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pages 990–998, 2008.
11. M. Welling, M. Rosen-Zvi, and G. E. Hinton. Exponential family harmoniums with an application to infomration retrieval. In *Proceedings of the 17th Neural Information Processing Systems (NIPS'05)*, 2005.
12. E. P. Xing, R. Yan, and A. G. Hauptmann. Mining associated text and images with dual-wing harmoniums. In *Proceedings of the 21st Conference on Uncertainty in Artificial Intelligence (UAI'05)*, pages 633–641, 2005.