# Topic Distributions over Links on Web

Jie Tang⋆, Jing Zhang⋆, Jeffrey Xu Yu‡, Zi Yang⋆, Keke Cai†, Rui Ma†, Li Zhang†, and Zhong Su†
⋆ *Department of Computer Science and Technology, Tsinghua University, China*
*{tangjie, zhangjing, yz}@keg.cs.tsinghua.edu.cn*
†*IBM, China Research Lab, Beijing, China*
*{caikeke, maruicrl, lizhang, suzhong}@cn.ibm.com*
‡*Department of Systems Engineering and Engineering Management, Chinese University of Hong Kong*
*yu@se.cuhk.edu.hk*

***Abstract*—It is well known that Web users create links with different intentions. However, a key question, which is not well studied, is how to categorize the links and how to quantify the strength of the influence of a web page on another if there is a link between the two linked web pages. In this paper, we focus on the problem of *link semantics analysis*, and propose a novel supervised learning approach to build a model, based on a training link-labeled and link-weighted graph where a link-label represents the category of a link and a link-weight represents the influence of one web page on the other in a link. Based on the model built, we categorize links and quantify the influence of web pages on the others in a large graph in the same application domain. We discuss our proposed approach, namely Pairwise Restricted Boltzmann Machines (PRBMs), and conduct extensive experimental studies to demonstrate the effectiveness of our approach using large real datasets.**

## I. Introduction

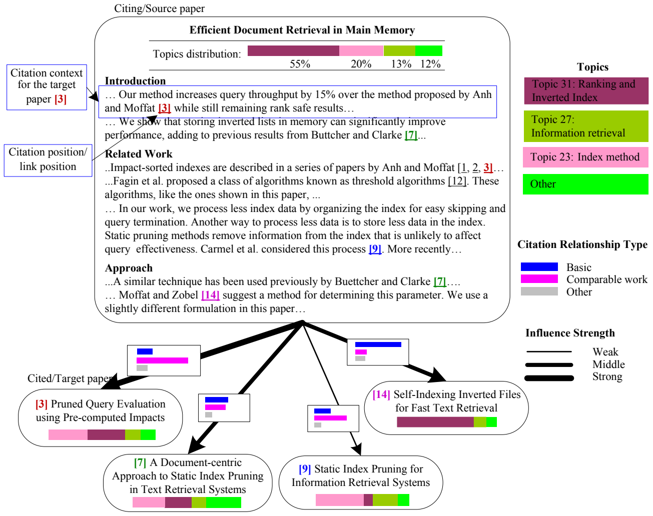
Web users create links with significantly different intentions. In the traditional Web, some links are created to help the reader navigate (drill down) to a web page with more detailed information about a terminology (or anchor text); some other links may point to previously published web pages with similar or even the same content; others may be only used to advertise a product. In social networking applications, people may create links to friends. Some of the links may point to family members; some others may connect to colleagues. Understanding of the category and the influence of each link can substantially benefit many applications.

There are a few works about link influence analysis. For example, Dietz et al. [1] propose a citation influence topic model to model the influential strength between papers. In social networks, several attempts have been made for analyzing the influential strength of the social relationship. For instance, [2] studies the correlation between social similarity and influence. [3] presents a method to quantify the influential strength. However, all of these works do not consider the category of every link.

To clearly motivate this work, we demonstrate with an example in a research paper citation context. As shown in Figure 1, there is a paper entitled "Efficient Document Retrieval in Main Memory" which addresses the efficient ranking and indexing method for information retrieval. We call it a source paper. The source paper cites 4 target papers, denoted as "[3]", "[7]", "[9]", and "[14]" due to various reasons. For example, as can be seen from the citation context words surrounding "[14]" in the source paper, the source paper adopts the method proposed in "[14]" as a basic component on top of which the work in the source paper is developed. As another example, the source paper cites "[3]" several times. The citation context of "[3]" in the source paper indicates that the method proposed in the source paper performs better than the work reported in the target paper "[3]". It gives a strong implication that the source paper and the target paper "[3]" address the similar problem as comparable work. The two relationships between the source paper and the target paper "[14]" and between the source paper and the target paper "[3]" are somehow different. The former indicates that "[14]" is a basic component to be cited, and the latter indicates that "[3]" is a comparable work.

After link semantics analysis, ideally we hope to provide users with the following information: (1) the topical aspects discussed in a web page; (2) the category of the link between two web pages; and (3) the influential strength of each link. Specifically, as the example in Figure 1, we first identify several topics, for example, efficient ranking (topic 31), indexing method (topic 23), and information retrieval (topic 27). Second, we determine the topic distribution for each paper, and such topics distribution may be different depending on whether the paper is a source paper or a target paper. Third, based on the learned topic distributions, we determine the category distributions on links and quantify such links. We refer to this problem as *link semantic analysis*.

In this paper, we aim to systematically investigate the link semantics analysis problem. The main contributions of the work are summarized below. First, we formalize our problem, and discuss multi-topic discovery, link category annotation, and link influence estimation. Second, we propose a unified model, called Pairwise Restricted Boltzmann Machines, which contains two layers of latent variables. One is for modeling topic distribution of documents, and the other is for modeling link categories and quantifying influences. Several learning algorithms including generative


Citing/Source paper
Efficient Document Retrieval in Main Memory
Topics distribution:
55%
20%
13%
12%
Citation context for the target paper [3]
Citation position/ link position
Topics
Topic 31: Ranking and Inverted Index
Topic 27: Information retrieval
Topic 23: Index method
Other
Citation Relationship Type
Basic
Comparable work
Other
Influence Strength
Weak
Middle
Strong
Cited/Target paper
[3] Pruned Query Evaluation using Pre-computed Impacts
[14] Self-Indexing Inverted Files for Fast Text Retrieval
[7] A Document-centric Approach to Static Index Pruning in Text Retrieval Systems
[9] Static Index Pruning for Information Retrieval Systems

**Introduction**

… Our method increases query throughput by 15% over the method proposed by Anh and Moffat [3] while still remaining rank safe results…

… We show that storing inverted lists in memory can significantly improve performance, adding to previous results from Buttcher and Clarke [7]...

**Related Work**

..Impact-sorted indexes are described in a series of papers by Anh and Moffat [1, 2, 3]…

...Fagin et al. proposed a class of algorithms known as threshold algorithms [12]. These algorithms, like the ones shown in this paper, ...

… In our work, we process less index data by organizing the index for easy skipping and query termination. Another way to process less data is to store less data in the index. Static pruning methods remove information from the index that is unlikely to affect query effectiveness. Carmel et al. considered this process [9]. More recently…

**Approach**

...A similar technique has been used previously by Buettcher and Clarke [7]….

… Moffat and Zobel [14] suggest a method for determining this parameter. We use a slightly different formulation in this paper…

Figure 1. Motivating example.

learning, discriminative learning, and hybrid learning are proposed. Third, we conduct extensive performance studies to evaluate our proposed approach using two different genres of large data sets, namely, Wikipedia data and publication data. We confirm that the proposed approach can effectively identify the link category and quantify influence.

## II. Problem Formulation

Consider a directed graph, $G = (V, E)$. Here, $V$ is a set of nodes (e.g., web pages), and $E$ is a set of links, where a link is an ordered pair $e = (d^s, d^t) \in E$ from the source page $d^s \in V$ to the target page $d^t \in V$. We focus on textual features in a web page, and represent a web page as a sequence of words. In brief, a web page $d$, is represented as a sequence of $N_d$ words, denoted as $\mathbf{w}_d$, where a word in $\mathbf{w}_d$ is chosen from a given vocabulary of size $W$. A single web page may contain multiple topics. For example, an introduction of "iPod touch" describes different topics including promotion, specification, etc. The semantics of a topic is represented by a collection of words that have high correlation with the topic. Given a set of topics and a set of words, a correlation matrix $\mathbf{U}$ can be constructed, where an element $U_{jk}$ in the matrix indicates the weight of the correlation between a word $w_k$ and a topic $z_j$. A *link context* $\mathbf{w}_e$ is a subsequence of words in the source page $d^s$ surrounding the *link position* in the source page. Note that $\mathbf{w}_e \subseteq \mathbf{w}_d$. A link context of a link is a fixed size window of words around the link position. In addition, we model the understanding of a link by *link category* $c_e$ for a link $e = (d^s, d^t) \in E$. The link category $c_e$ explains why a source page $d^s$ links to a target page $d^t$, and is taken from a set of given link categories which are given for a specific application domain. For example, for the Wikipedia data, we define three link categories: "drill down" (link pointing from a term to a description page), "similar" (link between two web pages with the similar content), and "other" (neither "drill down" nor "similar"). We assign link a weight, $f$, to quantify the *link influence* of the target page $d^t$, to be linked, to the source page $d^s$.

**Problem Statement**: In this paper, we focus on a problem of categorizing the link topics and quantifying the link influence, and study the problem as a supervised learning problem. Given a training dataset as a link-labeled and link-weighted graph $G = (V, E)$ where the label of a link is the link category of the link and the weight of a link is the link influence of the link in an application domain. We build up a model based on the training dataset, and categorize/quantify a large graph in the same application domain.

The main tasks of the problem are given below. (1) Multi-topic discovery: discover what topics a web page is about and estimate the link-category mixtures. (2) Link category annotation: unveil why a source page links to a target page, and identify the category distribution for each link. (3) Link influence estimation: estimate the influential strength of the target page on the source page.

## III. Our Approach

We can consider several baseline methods for the analysis. For multiple topic discovery, we can use the state-of-the-art topic models LDA [4] or RBM [5]. For link category annotation, we can use the multi-class SVM based on words appearing in the link context [6]. In addition, we can further incorporate the topic model learned from LDA or RBM as features into the SVM models for link category annotation. However, as the topic distribution and the link category are usually (sometimes even strongly) intertwined, the baseline methods, which learn the topic distribution and the link category in a cascaded way, are insufficient for link semantics analysis.

Our main idea is to formalize the link semantics analysis problem in a two-layer graphical model and propose a Pairwise Restricted Boltzmann Machines (PRBMs) to automatically identify the link category and link influence.

### A. *Pairwise Restricted Boltzmann Machines (PRBM)*

For ease of explanation, we first explain the PRBM model without considering the link influences, which will be further discussed in Section III-D. A PRBM model has the following components: a set of observable variables $\{\mathbf{w}^s, \mathbf{w}^t, \mathbf{w}_e, c_e\}$ and two layers of hidden variables $\{\mathbf{z}, \mathbf{h}\}$. In our problem the observable variables include words $\mathbf{w}^s = \{w_i^s\}_i$ in the source page, words $\mathbf{w}^t$ in the target page, link context words $\mathbf{w}_e$, and the link category $c_e$. The first (bottom) layer of hidden variables include topics $\mathbf{z}^s$ of the source page and topics $\mathbf{z}^t$ of the target page. The second (top) layer of hidden variables $\mathbf{h}$ is used to bridge the observable variables and the hidden topics. In this way, each link is characterized by different features including topic distributions and words $\mathbf{w}_e$ in link context. Figure 2 shows the graphical representation of the PRBM model. A PRBM

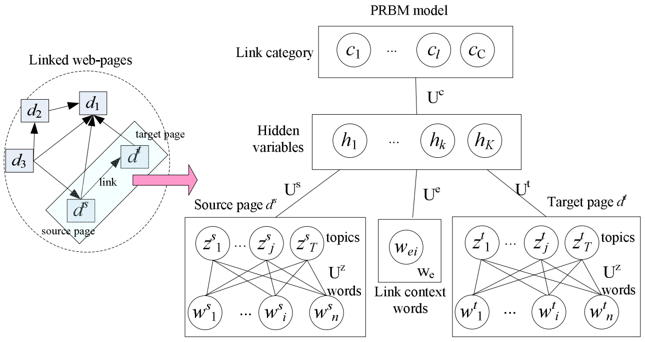


Figure 2. Graphical representation of PRBMs.

model is built for each link with its source page and target page. Correlation matrices $\{\mathbf{U}\}$ are defined between different types of variables. Specifically, $\mathbf{U}^z$ is the correlation matrix between topics and words in the web page. Each element in the matrix represents how likely a web page's content describes a topic. $\mathbf{U}^s$, $\mathbf{U}^t$, and $\mathbf{U}^e$ are correlation matrices to weigh different features. $c_i$ denotes the link category and $\mathbf{U}^c$ characterizes a category distribution of each link.

Given a labeled training dataset $D = \{(\mathbf{w}^s, \mathbf{w}^t, \mathbf{w})_i, c_i\}_1^N$, to train a PRBM model, we can consider finding a model that best fits the training data (with the maximal generative likelihood) while minimize the error of link categorization. However, it is non-trivial to train the PRBM model. Directly employing the traditional method, e.g., the training algorithm for RBM, is infeasible. In the PRBM model, a key difficulty is that a web page can be both the source page and the target page, thus a web page can have two different topic distributions. To solve this problem, we use a two-step algorithm for training the PRBM model. In the first step, words in all the web pages are used to learn a basic RBM model by maximizing the generative likelihood. The learned topic distribution for each web page is used as the initial value for the second step. In the second step, the initial topic distribution of each web page is used to train the second layer of latent variables $\mathbf{h}$, and, at the same time based on the initial topic distribution, two topic distributions are trained for each web page to represent its distribution of being source page and the distribution of being target page, by minimizing the categorization error or maximizing the generative likelihood of the link categories. The PRBM model thus makes use of the link information to update topic distributions of linked web pages.

In the first step, we learn the correlation matrix $\mathbf{U}^z$ by minimizing the negative log-likelihood $\mathcal{L}'_{gen}(D) = -\sum_{i=1}^{M} \log p(\mathbf{w}_d)$. This can be done using any methods for RBM such as Contrastive Divergence [7], brief Langevin [8], and Bethe approximation [9]. In the second step, we focus on how to learn the correlation matrices $\Theta = (\mathbf{U}^s, \mathbf{U}^t, \mathbf{U}^e, \mathbf{U}^c)$. Again, we consider minimization of the negative log-likelihood:

$$\mathcal{L}_{gen}(D) = -\sum_{e=1}^{N} \log\, p(\mathbf{w}_e, \mathbf{c}_e) \qquad (1)$$

By the property of Restrict Boltzmann Machine [5], the probability over a link context $\mathbf{w}_e$ can be defined as:

$$p(\mathbf{w}_e, \mathbf{c}_e) = \frac{1}{Z} \sum_{\mathbf{h}_e} \sum_{\mathbf{z}^s} \sum_{\mathbf{z}^t} \exp^{(-E(\mathbf{z}^s_e, \mathbf{z}^t_e, \mathbf{w}_e, \mathbf{c}_e, \mathbf{h}_e))} \qquad (2)$$

where $Z$ is a normalization factor; for a same web page its topic distributions $\mathbf{z}^s$ and $\mathbf{z}^t$ are initialized with the same value from the first step; the energy function $E(.)$ is:

$$E(\mathbf{z}^s_e, \mathbf{z}^t_e, \mathbf{w}_e, \mathbf{c}_e, \mathbf{h}_e) = -\sum_{j=1}^{T}\sum_{k=1}^{K} U^s_{jk} z^s_{ej} h_{ek}$$
$$-\sum_{j=1}^{T}\sum_{k=1}^{K} U^t_{jk} z^t_{ej} h_{ek} - \sum_{i=1}^{W}\sum_{k=1}^{K} U^e_{ik} w_{ei} h_{ek} - \sum_{l=1}^{C}\sum_{k=1}^{K} U^c_{lk} c_{el} h_{ek}$$
$$-\sum_{j=1}^{T} b_j z^s_{ej} - \sum_{j=1}^{T} b_j z^t_{ej} - \sum_{i=1}^{W} o_i w_{ei} - \sum_{k=1}^{K} g_k h_{ek} - \sum_{l=1}^{C} s_l c_{el} \qquad (3)$$

here $T$ is the number of topics; $K$ is number of hidden variables in the second layer; $U^e_{ik}$, for example, represents the symmetric interaction term between visible variable $w_i$ and hidden variable $h_k$; similar for the other matrices; $b_j$, $o_i$, $g_k$, and $s_l$ are bias terms.

Now, we consider for simplicity binary input variables. For example, for web page $d$, $w_{di} = 1$ represents that the web page $d$ contains the word $w_{di}$.

### *B. Model Learning*

**Generative Learning.** We learn the PRBM model by minimizing the negative log-likelihood (Eq. 1). Typically, this can be done by estimating the gradient of the model parameters. The exact gradient, for any parameter $\theta \in \Theta$ can be written as follows:

$$\frac{\partial L_{gen}(D)}{\partial \theta} = \mathbb{E}_{P_0}[\frac{\partial}{\partial \theta} E(D)] - \mathbb{E}_{P_M}[\frac{\partial}{\partial \theta} E(\hat{D})] \qquad (4)$$

where $\mathbb{E}_{P_0}[.]$ denotes an expectation with respect to the data distribution and $\mathbb{E}_{P_M}$ is an expectation with respect to the distribution defined by the model. The expectation $\mathbb{E}_{P_M}$ is intractable. We use Contrastive Divergence [7] to compute a stochastic approximation of this gradient. This approximation replaces the expectation $\mathbb{E}_{P_M}$ by samples generated after a limited number of Gibbs sampling iterations, with the sampler's states for the visible variables initialized at the training sample. The notation with a ˆ indicates a variable reconstructed after the Gibbs sampling. $E(D)$ is the energy function over the original data and $E(\hat{D})$ is the energy function over the data after $T$ step reconstruction by Gibbs sampling. Details are omitted due to space limitation.

**Discriminative learning.** In some applications, our interests may only lie in correctly predicting the link category based on topic distributions of the linked web pages. Thus we can explicitly emphasize that we want to infer one modality from other modalities, and perform discriminative learning of the PRBM model. The discriminative log-likelihood is:

$$\mathcal{L}_{dis}(D) = -\sum_{d=1}^{M} \ln\, p(\mathbf{w}_d) - \sum_{e=1}^{N} \log\, p(\mathbf{c}_e | \mathbf{z}^s_e, \mathbf{z}^t_e, \mathbf{w}_e) \qquad (5)$$

The update rule can be simplified in the discriminative learning, because we can only reconstruct (sample) the variable that we want to predict. For link category prediction, we can safely remove the steps for sampling topics of source pages, topics of target pages, and words in link contexts. Correspondingly, the gradient of the correlation matrices $\mathbf{U}^s$, $\mathbf{U}^t$, and $\mathbf{U}^e$ should be changed to:

$$\frac{\partial \mathcal{L}_{dis}(D)}{\partial U^s_{jk}} = \mathbb{E}_{P_0}[z^s_{ej} h_{ek}] - \mathbb{E}_{P_T}[z^s_{ej} \hat{h}_{ek}] \quad (6)$$

$$\frac{\partial \mathcal{L}_{dis}(D)}{\partial U^t_{jk}} = \mathbb{E}_{P_0}[z^t_{ej} h_{ek}] - \mathbb{E}_{P_T}[z^t_{ej} \hat{h}_{ek}] \quad (7)$$

$$\frac{\partial \mathcal{L}_{dis}(D)}{\partial U^e_{jk}} = \mathbb{E}_{P_0}[w_{ei} h_{ek}] - \mathbb{E}_{P_T}[w_{ei} \hat{h}_{ek}] \quad (8)$$

**Hybrid learning.** When the training data is sufficient, a discriminative learning usually results in a higher quality model than a counterpart generative model [10]. However, when the data is limited, the discriminative learning may underperform the generative counterpart. In our case, it is difficult to obtain a large number of labeled links. Thus a better way is to adopt a hybrid discriminative/generative learning by combining the respective training criteria:

$$\mathcal{L}_{hybrid}(D) = \alpha \mathcal{L}_{gen}(D) + \mathcal{L}_{dis}(D) \quad (9)$$

For estimation, we combine the updated parameters $\mathbf{U}^s$, $\mathbf{U}^t$, $\mathbf{U}^e$, $\mathbf{U}^c$ of the generative model and the discriminative model using a parameter $\alpha$.

### C. Link Category Annotation

The task of link category annotation is to predict the category of each link. We use a two-step algorithm to solve this problem. In the first step, for a given link, we calculate the topic distribution of the source page and that of the target page by:

$$p(z_j = 1|\mathbf{w}) = \sigma(\sum_i U^z_{ij} w_i + a_j) \quad (10)$$

where $a$ is a bias term learned in the above process.

In the second step, we use the mean field algorithm [11] to predict the category, according to Algorithm 1. Mean field is a variational approximation method for estimating a true distribution. For predicting link category, we use mean field to estimate the probability $p(c|\mathbf{h})$. Thus, finally, we obtain a category distribution $\{p(c|e)\}$ on each link $e$.

```
Input: the learned parameters $\mathbf{U}^s$, $\mathbf{U}^t$, $\mathbf{U}^e$, $\mathbf{U}^c$, $\mathbf{U}^z$
Estimate $\mathbf{z}^s$ and $\mathbf{z}^t$ using Eq. (10);
Initialize the expectation $\mathbb{E}[c]$ of the category randomly;
Initialize the expectation $\mathbb{E}^{old}[c]$ of the category as infinity;
while mean(abs($\mathbb{E}^{old}[c] - \mathbb{E}[c]$)) ¿ 1e-3 do
    $\mathbb{E}^{old}[c] = \mathbb{E}[c]$;
    Estimate the expectation of hidden variables $\mathbf{h}_e$;
    Estimate the expectation $\mathbb{E}[c]$;
end
```

**Algorithm 1**: Predicting via mean field.

### D. Link Influence Estimation

Another task for link semantics analysis is to estimate the link influence. We use two different methods to calculate the influential strength: one is category independent and the other is category dependent. In our experiments, we used the latter method.

In the category independent method, we use KL-divergence, a standard measure of the difference between two probability distributions, to estimate the negative similarity of the linked web pages. The basic idea is: if two papers describe a similar content (a small divergence between their topic distributions), then the target page may have a strong influence on the source page. Thus the influential strength $f$ of edge $e$ is defined as:

$$f = KL(\phi^{d^s} \| \phi^{d^t}) = \sum_{k=1}^{K} \phi^{d^s}_k \log \frac{\phi^{d^s}_k}{\phi^{d^t}_k} \quad (11)$$

where $\phi^{d^s}_k = p(z_k|\mathbf{w}_{d^s})$ is the normalized result of $z^s_{ek}$ associated with the source page $d^s$ and $\phi^{d^t}$ is the normalized results for the target page $d^t$.

The other way to measure the influential strength, i.e., the category dependent method, is to directly incorporate the influence as a variable node $f$ into the PRBM model. Specifically, we use a new correlation matrix $\mathbf{U}^f$ to quantitatively characterize the influence with the hidden variables $\mathbf{h}$. The influence $f$ of a link is generated from the hidden variables $\mathbf{h}$, according to a Gaussian distribution:

$$p(f|\mathbf{h}_e) = Gaussian(f|\sum_{k=1}^{K} U^f_k h_{ek} + r, 1) \quad (12)$$

where $\sum_{k=1}^{K} U^f_k h_{ek} + r$ is the mean of Gaussian, and the variance is 1.

## IV. Experimental Results

### A. Experimental Setting

**Data sets**: The experiments are conducted on two data sets: Wikipedia data and citation data. For the first data set, we collect $14,468$ "article" pages and $25,817$ links from Wikipedia [1]. All the pages are chosen from the "Computer science" category and links are "smart" links between these pages. 50 words before and after each link position are extracted as link context words. The publication data set contains 978,504 papers and 14,215,473 citation relationships, extracted from ArnetMiner.org [12].[2] Link (citation) positions are identified using regular expression (e.g., "[1]" or "(Smith, et al., 2007)") and again 50 words surrounding each citation position are extracted as citation context words. In a paper, several references may appear together at a same position (e.g., "[1, 2, 6]"). In such case, we create a link (citation) for each reference and create a same citation context for each of them.

[1] http://en.wikipedia.org

[2] http://arnetminer.org

Human annotators conduct annotation on link categories/influence. Specifically, every link is manually categorized into one of the three classes (i.e., “drill down”, “similar”, and “other”) and is assigned with an influential grade (i.e., “strong”, “middle”, and “weak”). A spec is created to guide the annotation process. For example, the assessment of link category is carried out in terms of (1) surrounding words of the link context (e.g., “The technology for statistical NLP comes mainly from machine learning and data mining”.) and (2) whether two web pages talk about a same/similar topic. The assessment of the influence is mainly based on (1) the content similarity, (2) whether ideas or contents of the target page have been used in the source page, and (3) the number of common links that the source page and the target page have. For disagreement in the annotation, we ask for some other annotators and finally determine the annotation by “majority voting”. We preprocess each web page/paper by (a) removing stopwords and numbers; (b) removing words that appear less than three times in the data set; and (c) downcasing the obtained words.

**Evaluation measures**: We first present the quantitative performance of the proposed approach and compare it with several baseline methods, and then perform several case studies to show the correlation between the link semantics and the learned topic distributions.

In the following experiments, we employ hybrid learning for the PRBM model. In PRBMs, model selection includes tuning best values for the learning rate, the number of the hidden variable, and the parameter $\alpha$ in the hybrid learning. We tune the parameters by cross-validation. In our experiments, we finally set both the low level topic number $T$ and the high level topic number $K$ as 500, as well as both iteration numbers for learning the low level RBM and the high level RBM as 3000, the learning rate $\lambda$ as 0.1, and, for hPRBM, $\alpha$ as 0.005.

### B. Quantitative Performance

**Model accuracy.** We test the different learning algorithms: generative learning, discriminative learning, and hybrid learning, for the PRBM model, which are denoted as gPRBM, dPRBM, and hPRBM respectively. For comparison purpose, we implement three baseline methods: SVM, SVM+LDA, and SVM+RBM. In SVM, we use link context words as features to learn a classification model and use the the model to predict the link category. In SVM+LDA, we use LDA [4] to learn topic distributions of web pages and further incorporate the learned topic distributions as features for predicting the link category. In SVM+RBM, we alternatively use RBM [5] to learn the topic distribution of each web page. In the experiments, we evaluate the performance of different approaches in terms of Precision, Recall, and F1-measure, and Accuracy [13].

Tables I lists the five-fold cross-validation results for link category annotation. The results indicate that our approach with the best performance (by hPRBM) significantly outperforms the baseline methods (+10.55% than SVM, +6.87% than SVM+LDA, and +8.81% than SVM+RBM in terms of F1-measure). We can also see that by combining the topic model (from LDA and RBM), the SVM based method also achieves an improvement (+1.74% by RBM and 3.68% in terms of F1-measure).

Table I
ACCURACY OF LINK CATEGORIZATION (%).

| Approach | Type | Precision | Recall | F1-measure | Accuracy |
|---|---|---|---|---|---|
| SVM | drill down | 64.09 | 53.60 | 58.24 | 88.50 |
| | similar | 77.66 | 83.34 | 80.33 | 76.59 |
| | other | 64.63 | 59.58 | 61.77 | 79.71 |
| | Avg. | 68.79 | 65.51 | 66.78 | 81.60 |
| SVM+LDA | drill down | 66.31 | 58.56 | 61.71 | 89.24 |
| | similar | 80.69 | 85.47 | 82.93 | 79.71 |
| | other | 69.58 | 64.36 | 66.75 | 82.42 |
| | Avg. | 72.19 | 69.46 | 70.46 | 83.79 |
| SVM+RBM | drill down | 68.63 | 53.0 | 59.66 | 89.40 |
| | similar | 76.47 | 91.74 | 83.36 | 78.89 |
| | other | 76.65 | 53.28 | 62.54 | 82.58 |
| | Avg. | 73.92 | 66.02 | 68.52 | 83.62 |
| gPRBM | drill down | 69.77 | 65.22 | 67.42 | 88.07 |
| | similar | 77.33 | 87.22 | 81.98 | 79.01 |
| | other | 78.00 | 60.94 | 68.42 | 85.19 |
| | Avg. | 75.03 | 71.12 | 72.61 | 84.09 |
| dPRBM | drill down | 96.00 | 44.44 | 61.54 | 95.88 |
| | similar | 76.85 | 96.89 | 85.71 | 78.60 |
| | other | 75.00 | 37.50 | 50.00 | 80.25 |
| | Avg. | **82.62** | 59.61 | 65.75 | 84.91 |
| hPRBM | drill down | 65.38 | 68.00 | 66.67 | 93.00 |
| | similar | 83.54 | 92.31 | 87.71 | 84.77 |
| | other | 88.14 | 69.33 | 77.61 | 87.65 |
| | Avg. | 79.02 | **76.55** | **77.33** | **88.48** |

We further evaluate the performance of link influence estimation. We define a baseline method based on cosine similarity using words’ TF*IDF scores as features. Such a method has been previously used for social influence analysis [2]. The estimated influence is a continuous value. To compare the continuous influence with the annotated discrete grade (i.e., strong, middle, and weak), we use three normal distributions to fit the estimated link influence from the data. Specifically, each normal distribution is responsible for a grade, with its mean calculated by the expectation of the link influence of that grade in the training data. We estimate which distribution has the highest probability to generate the influence and assigned the grade to the estimated influence. Finally, we use accuracy to evaluate the performance of link influence estimation by our approach and the baseline method. The accuracy of our approach is 85%, clearly outperforming that (79.0%) of the cosine similarity-based method.

### C. Topical Analysis

**Category-topic-pair mixture analysis.** Table II shows the correlation of topics and the citation relationship category. For each category, we show three hot topic-pairs. The last column lists the influential strength of the citing topic on the cited topic. The score was calculated using KL-divergence (Eq. 11), thus a lower score indicates a stronger influence.

Table II
CATEGORY-TOPIC-PAIR MIXTURE ANALYSIS.

| **Topics ($z^s$)<br>of Citing Papers** | **Topics ($z^t$)<br>of Cited Papers** | **Influential<br>Strength** |
|---|---|---|
| **Drill down (Basic theory)** | | |
| Topic 37<br>(Probabilistic model) | Topic 51<br>(Maximum entropy) | 5.05 |
| Topic 37<br>(Probabilistic model) | Topic 14<br>(Theory) | 4.38 |
| Topic 13<br>(Semantic Web) | Topic 26<br>(Ontology) | 4.36 |
| **Similar (Comparable work)** | | |
| Topic 7<br>(Recommendation) | Topic 7<br>(Recommendation) | 0.00 |
| Topic 21<br>(Frequent pattern learning) | Topic 21<br>(Frequent pattern learning) | 0.00 |
| Topic 26<br>(Ranking & Inverted Index) | Topic 26<br>(Ranking & Inverted Index) | 0.00 |
| **Other** | | |
| Topic 26<br>(Ranking & Inverted Index) | Topic 43<br>(Information retrieval) | 4.87 |
| Topic 26<br>(Ranking & Inverted Index) | Topic 48<br>(Web mining) | 4.24 |
| Topic 9<br>(Clustering) | Topic 18<br>(Natural Language Processing) | 4.69 |

Table III
THE CORRELATION BETWEEN TOPICS AND CITATION CATEGORY (%).

| **Topics of Cited Papers** | **Drill down**<br>(Basic Theory) | **Similar**<br>Comparable Work | **Other** |
|---|---|---|---|
| Topic 51<br>(Maximum entropy) | 47.6 | 33.4 | 19.0 |
| Topic 7<br>(Recommendation) | 8.5 | 72.4 | 19.1 |
| Topic 43<br>(Information retrieval) | 23.4 | 41.3 | 35.3 |

From Table II, we can see that a strong dependency exists between the relationship category and the topic distribution. For example, for the "Basic theory" category, the citing and the cited topics are often different from each other, while for the "Comparable work" category, the hot topics such as Topic 7, Topic 21, and Topic 26, have a clearly self-citing behavior. Accordingly, citations of the "Comparable work" category also have a stronger influential strength.

**Correlation between topics and category.** Table III shows three example topics and the category-distribution of their cited-relationships. We see that the topic "Maximum entropy" is more theoretical, accordingly about half citations (47.6%) to papers of this topic belong to "Basic theory". The "Recommendation" topic tends to be more application-oriented, thus citations to papers of this topic are mainly (72.4%) categorized as "Comparable work".

**Case study.** We give several examples from the citation data set. Table IV lists several example links (citation relationships) associated with each category. The last column shows the influence strength of the target paper on the source paper. Clearly we see, for the "similar" category, the influence of the target paper on the source paper is stronger (with a lower KL-divergence score) than the other two categories; while the influence in the "drill down" is stronger than the "other" category.

Table IV
EXAMPLE ANALYSIS OF LINK CATEGORY AND INFLUENCE.

| **Source Paper** | **Target Paper** | **Influence** |
|---|---|---|
| Drill Down (Basic theory) | | |
| Latent dirichlet allocation | A variational bayesian framework for graphical models | 2.36 |
| A sentiment-aware model for predicting sales | Probabilistic latent semantic analysis | 1.93 |
| Mining and summarizing customer reviews | Foundations of statistical natural language processing | 3.25 |
| Similar (Comparable work) | | |
| Constraint-driven clustering | Max-min d-cluster formation in wireless ad hoc networks | 2.00 |
| Hmms and coupled hmms for multi-channel eeg classification | Gaussian observation hidden markov models for eeg analysis | 0.51 |
| Scalable collaborative filtering using cluster-based smoothing | Evaluating collaborative filtering recommender systems | 1.51 |
| Other | | |
| Latent dirichlet allocation | Overview of the first text retrieval conference | 3.93 |
| Utility scoring of product reviews | Introduction to modern information retrieval | 3.16 |
| Sentiment analyzer: extracting sentiments about a given topic | Finding parts in very large corpora | 2.71 |

## V. CONCLUSION AND FUTURE WORK

In this paper, we investigate the problem of quantifying link semantics on the Web. We formalize the problem and propose a unified model, called Pairwise Restricted Boltzmann Machines, to solve this problem. Various learning algorithms, including generative learning, discriminative learning, and hybrid learning are studied for the PRBM model. Experiments on two different data sets show that our proposed approach is effective in mining link semantics.

## VI. *ACKNOWLEDGMENTS

The first author is supported by NSFC grant (No. 60703059), National High-tech R&D Program (No. 2009AA01Z138), Chinese Young Faculty Research fund (No. 20070003093), and a research award from IBM. The third author is supported by a grant of RGC (No. 419008), Hong Kong SAR, China.

## REFERENCES

[1] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *ICML'07*, 2007, pp. 233–240.

[2] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri, "Feedback effects between similarity and social influence in online communities," in *KDD'08*, 2008, pp. 160–168.

[3] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *SIGKDD'09*, 2009, pp. 807–816.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, 2003.

[5] P. Smolensky, "Information processing in dynamical systems: foundations of harmony theory," *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pp. 194–281, 1986.

[6] H. Nanba and M. Okumura, "Towards multi-paper summarization using reference information," in *IJCAI'99*, 1999, pp. 926–931.

[7] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Computation*, vol. 14, pp. 1771–1800, 2002.

[8] I. Murray and Z. Ghahramani, "Bayesian learning in undirected graphical models: approximate mcmc algorithms," in *UAI'04*, 2004, pp. 392–399.

[9] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *NIPS'01*, 2001, pp. 689–695.

[10] A. Y. Ng and M. I. Jordan, "On discriminative vs. generative classifiers: a comparison of logistic regression and naive bayes," in *NIPS'02*, 2002.

[11] E. P. Xing, M. I. Jordan, and S. Russell, "A generalized mean field algorithm for variational inference in exponential families," in *UAI'03*, 2003, pp. 583–591.

[12] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: Extraction and mining of academic social networks," in *KDD'08*, 2008, pp. 990–998.

[13] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval.* ACM Press, 1999.