Information Processing and Management 49 (2013) 1181–1193

Contents lists available at SciVerse ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A link-bridged topic model for cross-domain document classification

Pei Yang [a,c,*], Wei Gao [b], Qi Tan [a], Kam-Fai Wong [c]

[a] Department of Computer Science, South China University of Technology, Guangzhou, China
[b] Qatar Computing Research Institute, Qatar Foundation for Education, Science and Community Development, Doha, Qatar
[c] Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong

ARTICLE INFO

Article history:
Received 11 January 2012
Received in revised form 11 February 2013
Accepted 14 May 2013

Keywords:
Cross-domain
Document classification
Transfer learning
Auxiliary link network

ABSTRACT

Transfer learning utilizes labeled data available from some related domain (source domain) for achieving effective knowledge transformation to the target domain. However, most state-of-the-art cross-domain classification methods treat documents as plain text and ignore the hyperlink (or citation) relationship existing among the documents. In this paper, we propose a novel cross-domain document classification approach called Link-Bridged Topic model (LBT). LBT consists of two key steps. Firstly, LBT utilizes an auxiliary link network to discover the direct or indirect co-citation relationship among documents by embedding the background knowledge into a graph kernel. The mined co-citation relationship is leveraged to bridge the gap across different domains. Secondly, LBT simultaneously combines the content information and link structures into a unified latent topic model. The model is based on an assumption that the documents of source and target domains share some common topics from the point of view of both content information and link structure. By mapping both domains data into the latent topic spaces, LBT encodes the knowledge about domain commonality and difference as the shared topics with associated differential probabilities. The learned latent topics must be consistent with the source and target data, as well as content and link statistics. Then the shared topics act as the bridge to facilitate knowledge transfer from the source to the target domains. Experiments on different types of datasets show that our algorithm significantly improves the generalization performance of cross-domain document classification.

1. Introduction

Traditional machine learning approaches make a basic assumption that the training and test data should be drawn from the same feature space and follow the same distribution. In many real-world applications, however, this independent and identically distributed (i.i.d.) assumption does not hold. It has been extensively demonstrated in the literatures that traditional leaning models perform drastically worse when the i.i.d. assumption no longer holds (Dai, Yang, Xue, & Yu, 2007; Pan & Yang, 2010). In contrast, transfer learning allows the domains, distributions, and feature spaces used in training being different from those in testing. It utilizes labeled data available from some related (or source) domain in order to achieve effective knowledge transformation from it to the target domain, which plays an important role in the areas of machine learning and data mining. If done successfully, knowledge transfer would greatly improve the performance of learning by

* Corresponding author at: Department of Computer Science, South China University of Technology, Guangzhou, China. Tel.: +852 39438461; fax: +852 26035505.
E-mail addresses: yangpei@scut.edu.cn (P. Yang), wgao@qf.org.qa (W. Gao), tanqi@scut.edu.cn (Q. Tan), kfwong@se.cuhk.edu.hk (K.-F. Wong).

0306-4573/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.ipm.2013.05.002

avoiding tremendously expensive data annotation effort. Many examples in knowledge engineering justified that transfer learning can be generally beneficial for different applications, such as document classification (Sarinnapakorn & Kubat, 2007), sentiment classification (Blitzer, Dredze, & Pereira, 2007; Blitzer & Kakade, 2011), collaborative filtering (Pan, Xiang, & Liu Nathan, 2010), and Web search ranking (Gao, Cai, Wong, & Zhou, 2010).

With the prosperity of the Internet, more and more text document collections become available which contain rich textual contents that are interconnected via complex hyperlinks or citations, such as encyclopedia websites (e.g., Wikipedia), research paper archives (e.g., CiteSeer), and user-generated media (e.g., blogs and microblogs). Such kind of data is characterized as analogous structures where each article describes a topic (or concept), which contains a title, abstract, content and some references. A typical example is the Wikipedia page on Support Vector Machine.[1] Compared to the documents in traditional information management, these types of data contain links in addition to content. The hyperlinks (or citations) among articles capture their semantic relations and provide additional insights about their relationships.

Most state-of-the-art transfer learning algorithms for document classification treat documents as plain text and ignore the structure of links (or citations). However, link structures provide important information regarding the properties of documents and their relationships. Since the links imply the inter-dependence among the documents, the usual i.i.d. (i.e., independent and identically distributed) assumption of documents does not hold any more (Zhu, Yu, Chi, & Gong, 2007). From this point of view, the existing cross-domain document classification methods that ignore the link structure may fail to capture the dependency and would be unable to fully mine the common knowledge between different domains. It turns out that the transfer of cross-domain information could be seriously hindered due to the incompleteness of the mined common knowledge.

In this paper, we propose a novel approach to combine content and link information simultaneously for cross-domain document classification. The basic idea is that documents in different domains may share some common topics from the point of view of both content information and link structure, which could be used to mutually reinforce the identification of common topics, thus to enhance the classification knowledge across related but distinct domains. However, there are two essential problems that challenge the procedure of integrating link structures with the shared topics. First, the link data are usually very sparse and the common parts indirectly connected between domains cannot be fully discovered and utilized. For this reason, we utilize an auxiliary link network to strengthen the co-citation relationship among documents by embedding the background knowledge into a graph kernel. Our method cannot only enrich the document representation by reducing the data sparseness, but also enlarge the dimensional feature space by introducing new features shared by the documents, which would help fill the gap across domains. Secondly, it is difficult to come up with a unified model that combines the two types of information simultaneously because the learned decompositions of topics must be consistent with content and link statistics as well as the training and test data from different domains, following the basic principle of multi-view transfer learning regarding how to model the domain commonality and difference from the perspective of multiple views. To deal with this problem, we propose a probabilistic Link-Bridged Topic model (LBT) based on Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) for cross-domain knowledge transfer using a multi-view approach. LBT correlates the domain-specific features and encodes the domain commonality and distinction, as well as view consistency and difference, into the shared topics. Then the shared topics act as a bridge which helps knowledge transfer from the source to target domain. We derive the log-likelihood objective function for LBT and use EM algorithm for its optimization. Experimental results based on two types of datasets demonstrate that our method outperforms state-of-the-art baselines including the semi-supervised learning algorithm Transductive SVM (Joachims, 1999), the traditional multi-view algorithm Co-Training (Blum & Mitchell, 1998), the large-margin-based multi-view transfer learner MVTL-LM (Zhang, He, Liu, Si, & Lawrence, 2011) and the content-based transfer learning algorithm TPLSA (Xue, Dai, Yang, & Yu, 2008).

The rest of the paper is organized as follows: Section 2 reviews the related work; Section 3 presents the proposed LBT model for cross-domain document classification; Section 4 discusses the experiments and analyzes the results; finally, we conclude in Section 5 with discussions on future work.

2. Related work

2.1. Transfer learning from single view

Semi-supervised learning (Zhu & Goldberg, 2009) addresses the problem that the labeled data may be too few to build a robust classifier by leveraging a large amount of unlabeled data. Some popular semi-supervised learning models include self-training (Yarowsky, 1995), EM-based models (Brefeld & Scheffer, 2004), Co-Training and multi-view (Blum & Mitchell, 1998), graph-based methods (Joachims, 2003), transductive support vector machines (Joachims, 1999), and collective classification (Bilgic & Getoor, 2008). Nevertheless, most of them assume that the training and test data must be in the same feature space following the same distribution. In contrast, transfer learning allows the domains, distributions, and feature spaces used in training and testing to be different (Pan & Yang, 2010). Transfer learning was extensively studied in the machine learning community over the last decade. Its underlying assumption is that multiple tasks share certain structures, and therefore, the tasks can mutually benefit from these shared structures.

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

Most existing transfer learning methods can be classified into three categories: instance-transfer, feature-transfer and parameter-transfer. Instance-level approach (Dai et al., 2007; Xiang, Cao, Hu, & Yang, 2010) assumes that some training examples in the source domain are similar to the data in target domain which can be used to train the model for the target domain. Re-weighting and importance sampling are two major techniques in instance-transfer learning. The basic idea of feature-level transfer learning is to learn a "good" feature representation that is effective in bridging the divergence between domains. The major techniques in feature-transfer approach (Pan et al., 2010; Raina, Ng, & Koller, 2006; Wang, Domeniconi, & Hu, 2008; Zhong et al., 2009) are to transform the original feature space into another low or high dimensional feature space that reduces the domain distance. Parameter-transfer approach assumes that the source and target domains share some parameters or priors of their models. The objective of parameter-transfer approach (Dayanik, Lewis, Madigan, Menkov, & Genkin, 2006; Fujino, Ueda, & Nagata, 2010) is to discover the shared hyper-parameters or priors between domains. Our method is essentially a feature-based transfer approach by transforming the original feature space to a latent space, which captures the domain commonality as well as discrepancy.

Several approaches to transfer learning such as (Xue et al., 2008; Yang, Chen, Xue, Dai, & Yu, 2009; Zhuang et al., 2010) make use of PLSA (Hofmann, 1999). PLSA is a widely used probabilistic model, which provides solid statistical foundation. PLSA could be considered as a probabilistic implementation of latent semantic analysis (LSA) (Deerwester, Dumais, Furnas, Thomas, & Harshman, 1990). In this model, each document is considered as the convex combination of several topics, where these topics or latent semantic variables are obtained using the maximum-likelihood principle. An extension to PLSA was proposed in (Cohn & Hofmann, 2000) for identifying principal topics of document collection as well as authoritative documents within those topics, which incorporates the hyperlink connectivity in the PLSA model by using a joint probabilistic model for connectivity and content. The model treats all the documents as being from the same domain and the distribution difference is not taken into consideration. Likewise, Erosheva, Fienberg, and Lafferty (2004) adopted a mixed membership model for words and references in journal publications but treated membership scores as random Dirichlet realizations. Unlike Erosheva et al. (2004) which only uses the references information among the publications themselves, we utilize an auxiliary link network to mine the indirect co-citation relationship among the documents, which could help bridge the domains gap. Yang et al. (2009) present a new learning scenario, heterogeneous transfer learning, which improves learning performance when the data can be represented in different feature spaces and where no correspondence between data instances in these spaces is provided. They extend PLSA to help transfer the knowledge from social Web data, which have mixed feature representations. Xue et al. (2008) propose TPLSA to incorporate both labeled and unlabeled data. The hidden variables are used to bridge the documents in training and test domains, and learned under a joint probabilistic model. TPLSA is based on a simultaneous decomposition of the contingency tables associated with term occurrence knowledge in documents from both training and test domains, which identifies the principal topics of the training data as well as documents in the test data that support those topics. Zhuang et al. (2010) propose the Collaborative Dual-PLSA model to simultaneously capture both the domain distinction and commonality among multiple domains. The proposed model has two latent factors, i.e. word concept and document class.

Our work is closely related to TPLSA (Xue et al., 2008). Unlike TPLSA which only uses content as bridge, in our work, we focus on combining the content and link information to enhance the view consistency, which is a key issue for the success of multi-view transfer learning as discussed below. Furthermore, we make use of auxiliary link network to alleviate the data sparseness and help knowledge transfer across domains. These two key points make our proposed model distinctive from TPLSA (Xue et al., 2008). To the best of our knowledge, there is no existing study that focused on incorporating auxiliary link network for cross-domain document classification.

2.2. Transfer learning from multiple views

Our work is also related to multi-view learning, where observations are represented by multiple independent sets of features (Ruping & Scheffer, 2005). Blum and Mitchell (1998) introduced Co-Training. The idea is to train one learner on each view of the labeled examples and then to iteratively have each learner to label the unlabeled examples that receive the highest confidence. They proved that two independent yet compatible views can be used to learn a concept in the PAC (Valiant, 1984) framework based on few labeled and many unlabeled examples. Following the idea, many people extended the original Co-Training approach (Brefeld, 2004; Ghani, 2002; Nigam, McCallum, Thrun, & Mitchell, 2000).

Most research on multi-view learning is within a single domain, and multi-view transferring learning is not common. Our work can be considered as a case of multi-view transfer. Tur (2009) proposed a co-adaptation algorithm, which extends the Co-Training algorithm with model adaptation techniques. Co-adaptation makes the existing model adaptive using machine-labeled data with some weight tuned using a held-out set. Co-adaptation is designed for inductive transfer learning which assumes the target domain has a small amount of labeled data. Zhang et al. (2011) proposed a framework for multi-view transfer learning with a large margin approach. The labeled data from the source domain are weighted and used to construct a large margin classifier for target domain, and data from both domains are used to ensure the classification consistency between different views. The instance-level approach assumes that some similar source training examples can be identified and reused to train the target model. However, the performance of instance-based approach is generally poor since new target features lack support from source data (Blitzer et al., 2011). We focus on feature-level multi-view adaptation or transfer, where knowledge transformation takes place in the multiple transformed feature spaces simultaneously and complementarily.

3. Link-bridged topic model

Transfer learning generally aims to identify and exploit the shared common structures and properties among different domains for knowledge transformation. However, if cross-domain document classification methods only focus on the data on the source and target domains themselves, they may fail to capture the common parts among the domains that are indirectly connected. We observe that the indirect common co-citation relationship can be enhanced and mined with the help of an auxiliary link network. Furthermore, we combine the content information and co-citation relations using a unified probabilistic model based on PLSA (Hofmann, 1999) which maps the data from both domains to the latent topic space. Such mapping could correlate the domain-specific features via the shared topics. Then the domain commonality and difference are characterized by the shared but differential topics. In other words, the documents from both domains share some similar topics, while their associated probability distributions with the shared topics are to some extent discrepant. On the other hand, since the documents are described with multiple views (i.e., content and links), it can be expected that the generalization capacity of the model will be further enhanced by leveraging the complementary interactions between different views.

3.1. Problem statement

Suppose we are given the document collection in two related but different domains. Let D_s be the labeled documents set from the source domain, D_t be the unlabeled documents set from the target domain. Define $D = D_s \cup D_t$. The source and target data are assumed to draw from different feature spaces where the i.i.d. assumption no longer holds. Some features are defined in source or target domain only while some others are defined in both domains. For the easy of cross-domain feature transformation in the later stage, we technically expand the feature space in pre-processing to include all features from both domains into a unified space, where the missing features in either domain are replenished as 0.

Most machine learning algorithms use a feature vector as representation of instances. In this work, we will use two forms of feature-vector to represent the document, i.e. bag-of-words and bag-of-links, in order to incorporate the link structures. Let W be the vocabulary of the document collection. Each document d is represented by a bag-of-words set $\{w|w \in d \wedge w \in W\}$. Let $n(w, d)$ be the frequency of term w appearing in document d. Let C be the set of all the hyperlinks (or citations) in the collection. Each document d can be also represented by a bag-of-links set $\{c|c \in d \wedge c \in C\}$. Let $m(c, d)$ be the "frequency" of link c appearing in document d. Let Z be the topic label set. Each source training instance $d_s \in D_s$ is assigned with a unique topic label $z \in Z$. Let $f: D \rightarrow Z$ be a function to map the document $d \in D$ to a unique topic label $z \in Z$. Our objective is to assign a label $z \in Z$ to the unlabeled target document $d_t \in D_t$ as accurately as possible using the source training dataset.

3.2. Bridging domain gaps using auxiliary link network

A common problem for the link data is that they are normally very sparse. Considering the research papers, the average number of references in a paper is normally just about 20–30. Thus, how to alleviate the sparseness of link data is a key issue for building a robust cross-domain classifier using link structures among the documents.

On the other hand, there are some indirect correlations among the documents which cannot be directly discovered from the link information in the concerned document collection itself. We observe that such kind of indirect co-citation relationship can be enhanced and mined with the help of an auxiliary link network. For example, given the research papers in the areas of classification and clustering, we may turn to ACM paper citation network to find more indirect co-citation relationship among the articles from these two different domains, which would help alleviate the sparseness of link data. In this regard, the ACM paper network can be regarded as an example of "bigger world" which could provide extra useful background knowledge. A chain of, "a friend of a friend" statements can be made in an auxiliary network, to connect any two documents, which can bring two indirectly related documents closer together. The mined common citations can enrich the original link set and act as a bridge, which can be used to further fill the gap across domains and help the transfer.

We define the document set V_0 as the union of D and the external documents linked by documents in D. Given the citation relationship among the documents in V_0, we can construct the graph $G_0 = (V_0, E_0)$ where the vertex set V_0 represents documents and the edge set E_0 represents the hyperlinks (or citations) between documents. Note that the training and test dataset D is a subset of the vertex set V_0, i.e., $D \subseteq V_0$, and V_0 also contains those documents external to D but linked by documents from D. Let A_0 denote the adjacency matrix of G_0. To build up the bag-of-links vectors, the traditional way is to estimate the link frequency by using a fragment of the adjacency matrix A_0. As mentioned above, the data may be very sparse.

Next, we will illustrate how to leverage the auxiliary network to alleviate the data sparseness and facilitate cross-domain knowledge transfer. Suppose graph $G = (V, E)$ is an auxiliary network where the vertex set V represents documents in the auxiliary network and the edge set E represents the hyperlinks (or citations) between documents, and G_0 is a sub-graph of G. The advantage of incorporating the auxiliary network is that it can introduce more nodes and edges which are not included in the graph G_0 and provide more background knowledge among the documents. Let A denote the adjacency matrix of G. Then we define a base similarity matrix as follows:

$$B = A + A^T \tag{1}$$

where A^T is the transpose of A. Note that B is symmetric. We then use an exponential diffusion graph kernel (John & Nello, 2004) to mine the co-citation relationship as follows:

$$S = e^B = \sum_{k=0}^{\infty} \frac{1}{k!} B^k = \sum_{k=0}^{\infty} \frac{1}{k!} \left[A^k + (A^T)^k \right] \tag{2}$$

where k refers to the setting of the length of the reachable path between two nodes. Note that B reflects the direct citing and cited relationship between documents and B^k reflects the k-length-path indirect citing and cited relationship. Here we treat the citing and cited relationship indiscriminately. The exponential matrix e^B includes a decay factor $\frac{1}{k!}$ that we can use to control the contribution of longer paths.

The computation of Eq. (2) is intractable due to the infinite progression. In order to simplify the computation of S, we can rewrite it as another form. Since matrix B is symmetric, there exists an orthogonal matrix $U = [\xi_1, \ldots, \xi_l]$, where $\xi_i (1 \leqslant i \leqslant l)$ is the eigenvector of B, to diagonalize B as follows:

$$B = UDU^T \tag{3}$$

where $D = diag(\lambda_1, \ldots \lambda_l)$ is a diagonal matrix and $\lambda_i (1 \leqslant i \leqslant l)$ is the eigenvalue of B. Then we can obtain

$$S = e^B = \sum_{k=0}^{\infty} \frac{1}{k!} B^k = \sum_{k=0}^{\infty} \frac{1}{k!} UD^k U^T = Ue^D U^T \tag{4}$$

As a result, the calculation of Eq. (4) becomes much simplified since D is a diagonal matrix.

With the background knowledge introduced by the auxiliary network, the estimation of the link "frequency" $m(c, d)$ will become more accurate, which can be formulated as the following:

$$m(c, d) = S_{ij} (d = v_i \wedge c = v_j) \tag{5}$$

Intuitively, several reasons may account for why the co-citation mined from the auxiliary network would help to knowledge transfer between domains. Firstly, the indirectly related documents become correlated when the indirect co-citation relationships are taken into consideration by the graph kernel, and these common co-citation relationships can be enhanced and mined with the help of an auxiliary link network. It can be expected that the sparseness of link data can be significantly alleviated using the common co-citation relationships. Secondly, auxiliary network can introduce a number of new "common friends" shared by the documents which can be viewed as extra features. Then these extra features can be appended into the original feature space. Therefore, the gap between domains would be narrowed by mapping the documents from the original feature space to a higher dimensional feature space.

Fig. 1 shows an illustrated example. All the nodes in the graph refer to documents and the edges refer to the citation relationship between nodes. In order to clarify the different roles they play in the model, we use rectangle node to denote document and round node to link feature which is also a document. Here we have three document sets, i.e., $D = \{v_1, v_2\}$ where v_1 and v_2 are from the source and target domain, respectively, $V_0 = \{v_1, v_2, v_3, v_4, v_5\}$ and $V = \{v_1, v_2, v_3, v_4, v_5, v_6, v_7\}$. For simplicity, we use "small world" enclosed with dashed ellipse to represent the document set V_0, and "big world" enclosed with solid ellipse to represent the document set V. Note that $D \subseteq V_0 \subseteq V$. Let us firstly consider the small world. Given the graph, we can represent the documents as bag-of-links where the feature values are link frequencies, which are a fragment of adjacent matrix A_0, as shown in Fig. 2a. The two documents may not share any features and the document-link matrix is very sparse. In other word, there exists a gap between the source and target domains. However, the big world may provide more complete background knowledge that helps to bridge the domain gaps. Here the background knowledge is the direct or indirect co-citation relationship. Consider the two nodes v_1 and v_2, where both v_1 and v_2 cite v_7 and v_7 cites v_4, and likewise, both v_1 and v_2 are cited by v_6. Hence, v_6 and v_7 can be viewed as the bridge that brings v_1 or v_2 closer. This indicates that v_1 is to some extent related to v_2. It is reasonable to use such kind of direct or indirect co-citation to measure the similarity between nodes. As a toy example, we compute S using Equation (2) by setting the maximum length of path to be 2. Fig. 2b shows an enriched

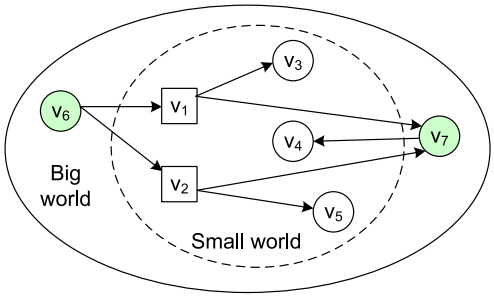

Fig. 1. Bridging domains using auxiliary link network.

	v_3	v_4	v_5
v_1	1	0	0
v_2	0	0	1

(a)

	v_3	v_4	v_5	v_6	v_7
v_1	1	0.5	0	1	1
v_2	0	0.5	1	1	1

(b)

Fig. 2. Represent document as bag-of-links vector. (a) The original feature vectors for the documents, which are sparse. (b) An enriched feature vectors by incorporating the background knowledge from an auxiliary link networ006B.

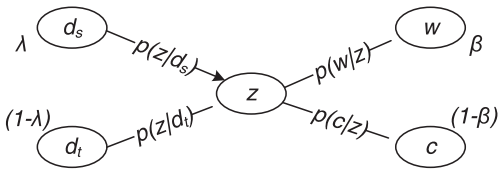

Fig. 3. Graphical model representation of LBT.

and enlarged document vectors which is a fragment of kernel matrix S. The feature value of v_4 for either v_1 and v_2 is 0.5, while it is 0 in Fig. 2a. v_6 and v_7 are extra features shared by v_1 and v_2. Obviously, v_1 is much closer to v_2 in Fig. 2b than in Fig. 2a.

3.3. The LBT model

The basic idea of our LBT model is to transform the original feature space to a latent space which captures the domain commonality as well as domain discrepancy. In this regard, LBT is a feature-transfer approach. Our model is based on PLSA (Hofmann, 1999). Intuitively, we can apply PLSA on the source and target data separately. However, since the source and target data are from related domains, they would share some similar topics. Therefore, it is advantageous to merge the two separate models into a joint probabilistic model which allows capturing the domain commonality, i.e., the shared topics by both domains. Meanwhile, since the domains are differential even though they are related, it is reasonable to associate the documents from both domains with the differential probabilities to the shared topics rather than with the equivalent probabilities. The rationale is that documents from both domains share some similar topics, while their associated probability distributions with the shared topics are to some extent discrepant. Such a merge-and-differentiate strategy enables the model to capture the domain commonality as well as domain difference.

At the meantime, our description of each document can be determined from two distinct views. The first is from the words occurring in this document, and the second is from the other documents with direct or indirect citation relationship with this document. The complementary interaction between the two views would help to discover more precise shared topics. Likewise, rather than applying two PLSA model on text and link data separately, it is beneficial to integrate them into a joint model.

Thus, in order to take into account all these factors mentioned above, we describe our model as a statistical generative process based on PLSA as follows, where the analogous notations from Hofmann (1999) are used:

- Firstly, select a document d_s with probability $p(d_s)$, or d_t with probability $p(d_t)$ from the documents set D.
- Secondly, pick a topic $z \in Z$ associated with d_s according to distribution $p(z|d_s)$, or the same associated with d_t according to distribution $p(z|d_t)$.
- Finally, given the topic z, generate a term w with distribution $p(w|z)$ and a link c with distribution $p(c|z)$.

Given the latent variables Z, we can define the following joint models, each consisting of a series of decompositions with regard to different topics:

$$p(w|d_s) = \sum_z p(w|z)p(z|d_s) \tag{6}$$

$$p(w|d_t) = \sum_z p(w|z)p(z|d_t) \tag{7}$$

$$p(c|d_s) = \sum_z p(c|z)p(z|d_s) \tag{8}$$

$$p(c|d_t) = \sum_z p(c|z)p(z|d_t) \tag{9}$$

The graphical representation of our Link-Bridge Topic (LBT) model is shown in Fig. 3. Note that in Eqs. (6) and (7) both decompositions of $p(w|d_s)$ and $p(w|d_t)$ share the same term-specific mixing component $p(w|z)$, and in Eq. (8) and (9) both decompositions of $p(c|d_s)$ and $p(c|d_t)$ share the same link-specific mixing component $p(c|z)$. Similarly, in Eq. (6) and (8) both decompositions of $p(w|d_s)$ and $p(c|d_s)$ share the same document-specific mixing component $p(z|d_s)$, and in Eq. (7) and (9) both decompositions of $p(w|d_t)$ and $p(c|d_t)$ share the same document-specific mixing component $p(z|d_t)$.

Specifically, the advantages of the proposed joint model are threefold:

- In our joint model, the expanded feature space of both domains are mapped into the latent topic space Z, which could correlate the seemingly unrelated source- and target-specific features by the topics if they have similar conditional probability $p(w|z)$ (or $p(c|z)$). Such correlation will then help bridge domain gap via the shared topics.
- On one hand, integrating the two separate PLSA models on source and target data into the joint model allows to capture the domain commonality, i.e., the shared topics by both domains. On the other hand, associating a topic $z \in Z$ with different distributions, $p(z|d_s)$ and $p(z|d_t)$, to the source and target data allows to capture the domain difference, i.e., the associated distributions with the topics are differential.
- Likewise, integrating the two separate PLSA models on text and link data into the joint model allows us to capture the view consistency, as well as their difference.

In summary, the power of our unified model is that it integrates all these kinds of correlations in a principled manner. The shared topics correlate the domain-specific features and encode the domain commonality and difference, as well as view consistency and difference. Since the mixing topics are shared, the learned decompositions must be consistent with the source and target data, as well as content and link statistics. As such, the shared topics act as the bridge to facilitate knowledge transfer from the source to the target domain.

Note that the description of each document can be generalized into multiple distinct views other than simply two. Therefore, our proposed model is further extensible by integrating multiple views of documents into this unified framework. Generally, we propose a novel method of combining multiple views of data in a latent topic model in order to transfer labels from one domain with labeled data to a different but related domain with unlabeled data via the shared latent topics. In this paper, we particularly focus on the case of documents with associated link network, but the proposed model is generally applicable to data with multiple views other than documents and links. In this regard, LBT can be viewed as a way of combining transfer learning and multi-view learning, which to the best of our knowledge has not received much attention in the literature.

3.3.1. How to transfer

Based on Fig. 3, we derive the log-likelihood objective function as follows:

$$L = \beta \sum_{w} \left[\lambda \sum_{d_s} \frac{n(w,d_s)}{\sum'_w n(w',d_s)} \log p(w|d_s) + (1-\lambda) \sum_{d_t} \frac{n(w,d_t)}{\sum'_w n(w',d_t)} \log p(w|d_t) \right] + (1 - \beta) \sum_{c} \left[\lambda \sum_{d_s} \frac{m(c,d_s)}{\sum'_c m(c',d_s)} \log p(c|d_s) + (1-\lambda) \sum_{d_t} \frac{m(c,d_t)}{\sum'_c m(c',d_t)} \log p(c|d_t) \right] \tag{10}$$

where $\frac{n(w,d)}{\sum'_w n(w',d)}$ and $\frac{m(c,d)}{\sum'_c m(c',d)}$ are normalization terms ensuring each document is given the same weight in the log-likelihood regardless of the number of observations associated with it, λ $(0 \leqslant \lambda \leqslant 1)$ acts as a tradeoff of weight between the training and test data, and larger λ indicates more reliance on the source training data set, β $(0 \leqslant \beta \leqslant 1)$ is the tradeoff between content and link, and larger β indicates content information is weighted more. When $\beta = 1$, the objective function ignores all the biases from link structure, and in this case, LBT model is equivalent to TPLSA (Xue et al., 2008); when $\beta = 0$, the objective function relies on link structure only, which is referred to as Link-bridged PLSA (LPLSA) in the rest of the paper. It is interesting to investigate how LPLSA using link structure only performs compared to TPLSA, which will be done in Section 4.Our goal is to maximize the log-likelihood L of the LBT model in Equation (10). Expectation–Maximization (EM) algorithm is used to find a local optimal solution of L.

- E-Step:

Given the term t and documents d_s and d_t, calculate the posterior probability of each topic z based on the old estimate of $p(w|z)$, $p(z|d_s)$ and $p(z|d_t)$:

$$p(z|w,d_s) = \frac{p(w|z)p(z|d_s)}{\sum'_z p(w|z')p(z'|d_s)} \tag{11}$$

$$p(z|w,d_t) = \frac{p(w|z)p(z|d_t)}{\sum'_z p(w|z')p(z'|d_t)} \tag{12}$$

Similarly, given the link c and documents d_s and d_t, calculate the posterior probability of each topic z based on the old estimate of $p(c|z)$, $p(z|d_s)$ and $p(z|d_t)$:

$$p(z|c,d_s)=\frac{p(c|z)p(z|d_s)}{\sum_z' p(c|z')p(z'|d_s)} \tag{13}$$

$$p(z|c,d_t)=\frac{p(c|z)p(z|d_t)}{\sum_z' p(c|z')p(z'|d_t)} \tag{14}$$

- M-step:

Given the posterior probability of each topic z, re-estimate conditional probability $p(w|z)$, $p(c|z)$, $p(z|d_s)$ and $p(z|d_t)$. Each of the below conditional probability is a mixture component of posterior probability of latent topics.

$$p(w|z)\propto\lambda\sum_{d_s}\frac{n(w,d_s)}{\sum_w' n(w',d_s)}p(z|w,d_s)+(1-\lambda)\sum_{d_t}\frac{n(w,d_t)}{\sum_w' n(w',d_t)}p(z|w,d_t) \tag{15}$$

$$p(c|z)\propto\lambda\sum_{d_s}\frac{m(c,d_s)}{\sum_c' m(c',d_s)}p(z|c,d_s)+(1-\lambda)\sum_{d_t}\frac{m(c,d_t)}{\sum_c' m(c',d_t)}p(z|c,d_t) \tag{16}$$

$$p(z|d_s)\propto\beta\sum_{w}\frac{n(w,d_s)}{\sum_w' n(w',d_s)}p(z|w,d_s)+(1-\beta)\sum_{c}\frac{m(c,d_t)}{\sum_c' m(c',d_t)}p(z|c,d_t) \tag{17}$$

$$p(z|d_t)\propto\beta\sum_{w}\frac{n(w,d_t)}{\sum_w' n(w',d_t)}p(z|w,d_t)+(1-\beta)\sum_{c}\frac{m(c,d_t)}{\sum_c' m(c',d_t)}p(z|c,d_t) \tag{18}$$

3.3.2. Algorithm for LBT

From the above equations, we can derive our LBT algorithm as shown in Algorithm 1. In the initial phase, the conditional probability $p(w|z)$ and $p(c|z)$ is set to be uniform. In order to utilize the label knowledge in the source training data, we initialize the conditional probability $p(z|d_s)$ for each labeled document $d_s \in D_s$ as follows:

$$p(z|d_s)=\begin{cases}\eta_s, & \text{if } f(d_s)=z\\ \frac{1-\eta_s}{|Z|-1}, & \text{otherwise}\end{cases} \tag{19}$$

where η_s is a randomly generated number, which is empirically required to meet the condition $0.5<\eta_s<1.0$. The idea is that the conditional probability $p(z|d_s)$ will receive a high value when the topic label z is the same with $f(d_s)$. For a unlabeled target document $d_t \in D_t$, the conditional probability $p(z|d_t)$ is set to be uniform since no prior knowledge is currently available.

Then the algorithm iteratively performs the E-Step and the M-Step in order to seek optimal log-likelihood based on objective function L in Equation (10). When it is converged, a unique topic label is assigned to the target documents according to $f(d_t)=\arg\max_z p(z|d_t)$.

Algorithm 1. Link-Bridged Topic Model

Input: Document-term matrices $D_s \times W$ and $D_t \times W$, document-link matrices $D_s \times C$ and $D_t \times C$. A topic label $z \in Z$ is assigned for each source document $d_s \in D_s$.

Output: Topic label $z \in Z$ assigned to each unlabeled target document $d_t \in D_t$.

1: Initialize conditional probability $p(w|z)$, $p(c|z)$, $p(z|d_s)$ and $p(z|d_t)$.
2: **WHILE** the change of L in Equation (10) between two sequential iterations is greater than a pre-defined threshold **DO**
3: **E-Step**: Update posterior probability $p(z|w,d_s)$, $p(z|w,d_t)$, $p(z|c,d_s)$ and $p(z|c,d_t)$ based on Eqs. (11)–(14) respectively.
4: **M-Step**: Re-estimate conditional probability $p(w|z)$, $p(c|z)$, $p(z|d_s)$ and $p(z|d_t)$ based on Eqs. (15)–(18) respectively.
5: **END WHILE**
6: **FOR** each unlabeled target document $d_t \in D_t$ **DO**
7: $f(d_t)$ = argmax$_z$ $p(z|d_t)$
8: **ENG FOR**

4. Experiments

In this section, we experimentally evaluate the proposed LBT algorithm for cross-domain document classification in comparison with the state-of-the-art algorithms as baselines. Two types of datasets, i.e., scientific research papers dataset and web pages dataset, are used for the evaluation.

4.1. Datasets and setup

Cora (McCallum, Nigam, Rennie, & Seymore, 2000) is an online archive of computer science research papers which contains approximately 37,000 papers, and over 1 million links among roughly 200,000 distinct documents. The documents in the dataset are categorized into a hierarchical structure. We select a subset of Cora papers for our model training and test, which contained five top-categories and 10 corresponding sub-categories (the numbers are in the parenthesis):

- DA_1="/data_structures__algorithms_and_theory/computational_complexity/" (711);
- DA_2="/data_structures__algorithms_and_theory/computational_geometry/" (459);
- EC_1="/encryption_and_compression/encryption/" (534);
- EC_2="/encryption_and_compression/compression/" (530);
- NT_1="/networking/protocols/" (743);
- NT_2="/networking/routing/" (477);
- OS_1="/operating_systems/realtime/" (595);
- OS_2="/operating_systems/memory_management/" (1102);
- ML_1="/machine_learning/probabilistic_methods/" (687);
- ML_2="/machine_learning/genetic_algorithms/" (670).

Note that each top-category contains several sub-categories, while we only select two sub-categories from each top-category to generate our datasets. Based on this data, we used a way similar to Pan and Yang (2010) to construct our training and test sets. For each set, we chose two top categories, one as positive class and the other as the negative. Different sub-categories were regarded as different domains. The task is defined as top category classification. For example, the dataset denoted as DA-EC consists of source domain: DA_1(+), EC_1(−); and target domain: DA_2(+), EC_2(−). The method ensures the domains of labeled and unlabeled data are related due to same top categories, but the domain distributions are different because they are drawn from different sub-categories. Such a preprocessing is a common practice for data preparation for adaptation purpose. The domain difference can be justified like some previous works (Pan & Yang, 2010) where it was found that SVM classifier trained on in-domain data performed much worse out of domain, which implies large domain gap.

The second dataset we use is the Industry Sectors dataset[2] which is a collection of about ten thousand Web pages belonging to companies from various economic sectors. The corporate Web pages are classified into a hierarchical structure. We chose a subset of Web pages from the five top sectors, i.e., energy, financial, healthcare, transportation and consumer and 10 corresponding sub-categories. Based on these five top-categories, we generated 10 datasets in a similar way to what we had done for the Cora datasets to ensure the domain relatedness as well as difference.

We preprocessed the data for both text and link information. For the texts, we removed stop words and low-frequency words with count less than 5. For the links, we removed the links with less than three citation counts. Then the standard TF-IDF (Salton & Buckley, 1988) technique was applied to both the text and link datasets.

4.2. Effectiveness of auxiliary network

Here we examine whether the embedding of the co-citation relationships mined from the auxiliary network into a graph kernel would lead to a better representation of the documents. For each original link dataset, the co-citation relationships among the documents from the two corresponding top-categories are used to construct the auxiliary link network. Then we employed such an auxiliary link network to generate the enriched link dataset (see Section 3.1). Finally, since the auxiliary network would introduce noisy link features, here we employ a simple feature selection mechanism which removes the features whose document frequencies are less than 3 in the dataset. We found this mechanism, though simple, worked well on the Cora datasets.

Since SVM (Joachims, 1999) has shown state-of-the-art performance compared to most of other supervised machine learning methods, we fed two kinds of link datasets to the SVM classifier for performance comparison:

- SVM-OL: SVM is applied on the original link dataset.
- SVM-L: SVM is applied on the enriched link dataset.

The classification error rate is used to evaluate the classification performance, which is defined as the number ratio between the misclassified test instances and the total test instances.

Table 1 shows the error rate on the Cora and the Sectors datasets. The bold items in Table 1 indicate the best results achieved by the algorithms for each dataset. For the Cora datasets, SVM-L significantly outperformed SVM-OL on most datasets. On average, the error rate of SVM-L is 23.0% lower than that of SVM-OL. It verifies that the auxiliary network can provide more complete background knowledge about the correlation among the documents which would help to reduce the domain gap. For the Sectors datasets, the performance superiority of SVM-L over SVM-OL is not significant. In comparison with the

[2] http://people.cs.umass.edu/~mccallum/data.html.

Table 1
Error rate for original and enriched datasets.

Cora datasets	SVM-OL	SVM-L	Sectors datasets	SVM-OL	SVM-L
DA-EC	0.5006	**0.4305**	Energy-Financial	0.485	**0.460**
DA-NT	0.2577	**0.1188**	Energy-Healthcare	0.475	**0.465**
DA-OS	0.5342	**0.0552**	Energy-Transportation	**0.498**	0.517
DA-ML	0.3588	**0.2245**	Energy-Consumer	0.495	**0.453**
EC-NT	0.5098	**0.3045**	Financial-Healthcare	0.490	**0.440**
EC-OS	0.3126	**0.2670**	Financial-Transportation	**0.482**	0.492
EC-ML	**0.4402**	0.6086	Financial-Consumer	0.480	**0.455**
NT-OS	**0.3920**	0.4781	Healthcare-Transportation	0.503	**0.442**
NT-ML	0.3265	**0.3255**	Healthcare-Consumer	**0.490**	0.521
OS-ML	0.2348	**0.1639**	Transportation-Consumer	0.492	**0.473**
Average	0.3867	**0.2977**	**Average**	0.489	**0.472**

Table 2
Error rates for the Cora datasets.

Cora datasets	TSVM-C	TSVM-L	TSVM-CL	Co-Training	MVTL-LM	TPLSA	LPLSA	LBT
DA-EC	0.286	0.168	0.130	0.145	0.176	0.155	0.084	**0.064**
DA-NT	0.179	0.119	0.076	0.099	0.097	0.091	0.122	**0.062**
DA-OS	0.271	0.257	0.243	0.106	0.064	0.048	0.036	**0.011**
DA-ML	0.205	0.109	0.104	0.125	0.159	0.087	0.045	**0.020**
EC-NT	0.301	**0.135**	0.155	0.224	0.213	0.189	0.188	0.157
EC-OS	0.360	0.790	0.201	0.091	0.161	0.106	0.051	**0.044**
EC-ML	0.338	0.216	0.175	0.219	**0.132**	0.188	0.790	0.157
NT-OS	0.372	0.552	0.471	0.460	0.275	0.102	0.129	**0.045**
NT-ML	0.223	0.107	0.102	0.138	0.067	0.051	0.061	**0.022**
OS-ML	0.211	0.335	0.159	0.037	0.119	0.053	0.015	**0.006**
Average	0.275	0.279	0.182	0.164	0.146	0.107	0.152	**0.059**

research papers, the Web pages would contain much more noisy links. In this case, the import of auxiliary network not only provides background knowledge, but also brings with more noisy data, such as advertisement links. Next we will explore a more effective feature selection mechanism to filter out the noisy data in the Web pages.

4.3. Algorithms comparison and analysis

We compared our model with state-of-the-art algorithms including the semi-supervised learning method Transductive SVM (Joachims, 1999), the traditional multi-view algorithm Co-Training (Blum & Mitchell, 1998), the large-margin-based multi-view transfer learner MVTL-LM (Zhang et al., 2011) and the content-based transfer learning algorithm TPLSA (Xue et al., 2008). We also compared our LPLSA (see Section 3.2) with TPLSA. Note that LPLSA is a special case of our proposed model.

For simplicity, we used the postfix *-C*, *-L* and *-CL* to denote that the classifier was fed with the text, link and merged dataset, respectively. All the link-based datasets used in the following experiments refer to the enriched link datasets. Both the text and link datasets were fed to the multi-view classifiers Co-Training, MVTL-LM and LBT. TPLSA relies on text information only, thus we applied TPLSA on content-based datasets. LPLSA is fed with the link-based datasets. The comparison results are shown in Tables 2 and 3, where the bold items indicate the best results achieved by the algorithms for each dataset.

Table 2 shows the error rate on the Cora datasets. TSVM performed poorly for adaptation when using either content or link features. Simply merging the two sets of features make some improvements, implying that text and link can be complementary, but it may degrade the confidence of classifier on some instances whose features become conflict because of merge. Co-Training can avoid this problem by boosting the confidence of classifiers built on the distinct views in a complementary way, thus performing better than TSVMs. Since both TSVM and Co-Training don't consider the distribution gap, they performed clearly worse than the transfer learning based approaches including MVTL-LM, TPLSA, LPLSA, and LBT.

Among the twenty datasets, although generally a little worse, LPLSA achieves comparable performance on most data sets compared to TPLSA. This demonstrates that link structure reflecting the inter-dependence relationship among the documents is of great value for cross-domain document classification. However, we find that TPLSA significantly outperforms LPLSA on the dataset EC-ML. One possible reason for this is that the papers under the three sub-categories, i.e., EC_1 (/encryption_and_compression/encryption/), EC_2 (encryption_and_compression/compression/) and ML_1 (machine_learning/probabilistic_methods/), may share co-citation relationship with those papers of some common topics, such as information theory, rendering it difficult for LPLSA to distinguish ML_1 from EC_1 and EC_2.

As shown in Table 2, LBT outperforms TPLSA significantly on all data sets. Two reasons may account for the advantage of LBT over TPLSA and LPLSA. Firstly, LBT employs the auxiliary link network to discover more shared co-citation between domains which helps alleviate the data sparseness and leads to a better representation of documents. Link structure provides importance information about the relationship among documents. Secondly, LBT exploits both the content information and

Table 3
Error rates for the Sectors datasets.

Sectors datasets	TSVM-C	TSVM-L	TSVM-CL	Co-Training	MVTL-LM	TPLSA	LPLSA	LBT
Energy-Financial	**0.109**	0.480	0.119	0.223	0.254	0.287	0.460	0.188
Energy-Healthcare	0.089	0.455	**0.050**	0.252	0.198	0.158	0.332	0.134
Energy-Transportation	0.343	0.483	0.393	0.408	0.386	0.403	0.423	**0.284**
Energy-Consumer	**0.114**	0.448	0.240	0.287	0.327	0.342	0.450	0.292
Financial-Healthcare	0.290	0.435	0.275	0.335	0.314	0.210	0.425	**0.120**
Financial-Transportation	0.091	0.467	**0.010**	0.357	0.276	0.176	0.347	0.136
Financial-Consumer	0.480	0.420	0.400	0.385	0.334	0.315	0.395	**0.290**
Healthcare-Transportation	0.126	0.477	**0.040**	0.146	0.167	0.111	0.452	0.110
Healthcare-Consumer	0.370	0.510	0.415	0.275	0.251	**0.190**	0.510	0.195
Transportation-Consumer	**0.130**	0.467	0.150	0.402	0.267	0.231	0.452	0.216
Average	0.214	0.464	0.209	0.307	0.277	0.242	0.425	**0.196**

link structure that are incorporated into a unified link-bridged topic model. The shared topics play a key role in the knowledge transfer from source domain to target domain. Moreover, the LBT classifier can be regarded as a tradeoff between the content-based classifier and the link-based classifier. The complementary cooperation between two types of classifiers would help discover more precise shared topics and thus commonly yield better prediction performance.

It is shown that multi-view adaptation using MVTL-LM performed worse than LBT on most datasets. Generally, it suggests that instance-based approach relying on instance weighting is not effective when the data of different domains are drawn from different feature spaces. Although MVTL-LM regulates view consistency on both domains' instances, it cannot identify the useful association between target-specific and source-specific features, which is the key to the success of adaptation especially when domain gap is large and less commonality could be found. In contrast, LBT uses a unified probabilistic model to find such correlations. Such common knowledge smoothly bridges the gap between different domains and enhances the generalization ability of the cross-domain classifiers.

Table 3 shows the error rate on the Sectors datasets. We can see that TSVM-L and LPLSA performed significantly worse than TSVM-C and TPLSA, respectively. This is because the link data contain much noise which seriously deteriorates the prediction performance. On average, TSVM-CL performed a little better than TSVM-C. Likewise, LBT outperform TPLSA on most datasets. It indicated that though the link data is noisy and sparse, the two views are still complementary, which helps improve the prediction performance.

On most datasets, LBT significantly outperformed Co-Training and MVTL-LM. However, the advantage of LBT over TSVM-CL on the Sectors datasets is not comparable as it is on the Cora datasets. It suggested that our proposed approach is more effective on the research paper datasets than on the Web page datasets. Intuitively, unlike the scientific papers, the Web pages contain much noisy data, which likely lead to topic drift. Since our proposed method is based on the probabilistic topic model, it would suffer from the topic drift problem.

4.4. Parameter sensitivity

Here we aim to study how the parameters, such as λ and β, affect the performance of the proposed algorithm. The results are shown in Figs. 4 and 5. Fig. 4 shows the error rate curves for different β across different Cora data sets. The parameter β $(0 \leqslant \beta \leqslant 1)$ acts as a tradeoff of weight between content and link. Note that larger β indicates more reliance on the content information of documents. In this figure, the X-axis shows the change of parameter β which varies from 0.0 to 1.0. The Y-axis represents the error rate of classification across different datasets. As shown in Fig. 4, the error rate will first decrease and then increase when β increases. The algorithm performs worse nearly on all of the data sets when heavily relying on either the content information $(0.9 \leqslant \beta \leqslant 1.0)$ or the link structure $(0 \leqslant \beta \leqslant 0.1)$. It is shown that setting β at the interval from 0.5 to 0.8 will achieve the best performances for LBT across most of the datasets. It suggests that the two views of document are complementary and beneficial to the classifier.

Fig. 5 shows the error rate curve for different λ across different Cora data sets. The parameter λ $(0 \leqslant \lambda \leqslant 1)$ acts as a tradeoff of weight between training and test data. Larger λ indicates heavier reliance on the source training data. In this figure, the X-axis shows the change of parameter λ which is varied from 0.1 to 1.0. It is shown that setting λ at the interval from 0.4 to 0.7 will achieve higher performance for LBT across most of the datasets. But there are some exceptions. For example, the algorithm performs better on the EC-NT dataset when setting λ at the interval from 0.7 to 0.9. In most cases, the algorithm performs worse when relying only on the source training set $(\lambda = 1)$. It suggests that the distribution between the training and test data from different domains are different, and the algorithm can perform better by taking such kind of distribution difference into consideration than by treating them indiscriminately.

As a result, we tuned the parameters β and λ, by using cross-validation on the training dataset for all the experiments.

4.5. Convergence

We tested the convergence property of LBT as well. Fig. 6 shows the experimental results. The X-axis represents the number of iterations. The Y-axis, ΔL, represents the change of likelihood L in Eq. (7) between two sequential iterations across

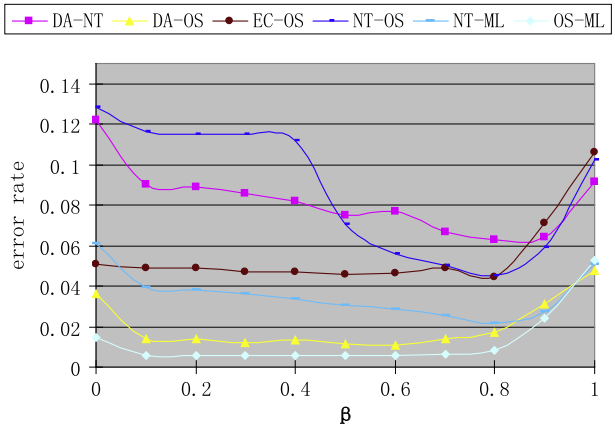

Fig. 4. Error rate curve for different β.

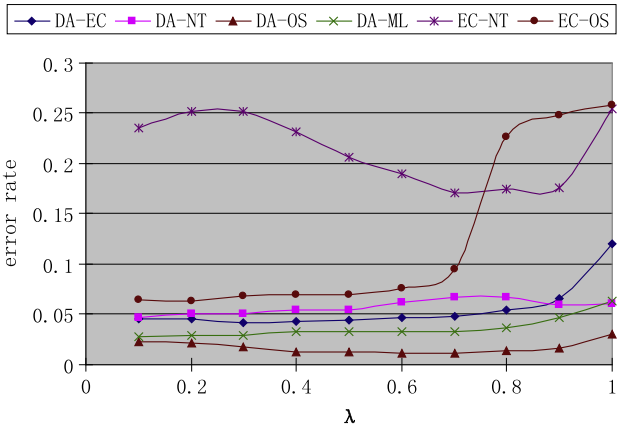

Fig. 5. Error rate curve for different λ.

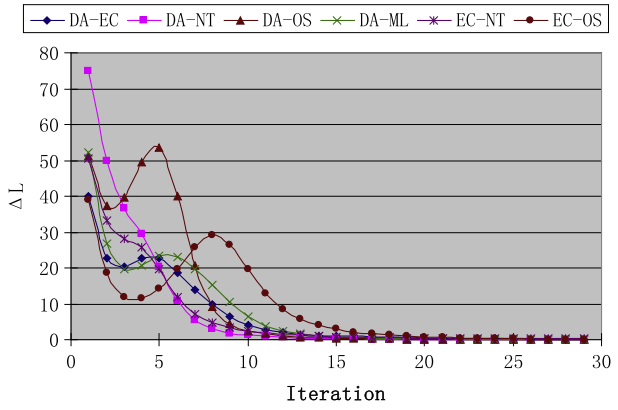

Fig. 6. ΔL for different number of iterations.

different datasets. LBT uses EM algorithm to find a local optimal point. The EM algorithm performs the E-step and M-step iteratively, and the convergence is guaranteed. As shown in Fig. 6, the change of likelihood L decreases very fast during the first 15 iterations and becomes stable after 30 iterations. Thus, we terminated the algorithm after a maximum of 30 iterations.

5. Conclusion

We introduce a novel method called Link-Bridged Topic (LBT) model for cross-domain document classification. Firstly, we employ the auxiliary link network to discover the shared co-citation relationship between documents in different domains. Then we combine the content information and link structures among documents using a unified probabilistic model to mine the hidden common topics. Based on the sharing structure, the LBT model achieves effective knowledge transformation between different domains. The experimental results demonstrate that compared to the state-of-the-art baseline algorithms our algorithm significantly improves the prediction accuracy of cross-domain document classification.

Embedding the background knowledge mined from the auxiliary link network into a graph kernel can enhance the co-citation relationships among documents and thus help knowledge transfer between domains. However, auxiliary network would also introduce the noisy data. Next we will design a more refined feature selection mechanism to filter out noise data.

Also, transfer learning would hurt the performance when the domains or multi-views are too dissimilar. As part of ongoing work we are exploring the boundary between positive transfer and negative transfer and learning how to measure the extent of relatedness between domains and multi-views.

References

Bilgic, M., & Getoor, L. (2008). Effective label acquisition for collective classification. In *KDD-2008* (pp 43–51).

Blitzer, J., Dredze, M., & Pereira, F. (2007). Biographies, bollywood, boomboxesand blenders: Domain adaptation for sentiment classification. In *ACL-2007* (pp. 432–439).

Blitzer, J., Kakade, S., & Foster Dean, P. (2011). Domain adaptation with coupled subspaces. In *AISTATS-2011* (pp. 173–181).

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT-1998* (pp. 92–100).

Brefeld, U., & Scheffer, T. (2004). Co-EM support vector learning. In *ICML-2004*.

Cohn, D. A., & Hofmann, T. (2000). The missing link – A probabilistic model of document content and hypertext connectivity. In *NIPS-2000* (pp. 430–436).

Dai, W. Y., Yang, Q., Xue, G. R., & Yu, Y. (2007). Boosting for transfer learning. In *ICML-2007* (pp. 193–200).

Dayanik, A. A., Lewis, D. D., Madigan, D., Menkov, V., & Genkin, A. (2006). Constructing informative prior distributions from domain knowledge in text classification. In *SIGIR-2006* (pp. 493–500).

Deerwester, S., Dumais, S. T., Furnas, G. W., Thomas, K. L., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 63*, 391–407.

Erosheva, E. A., Fienberg, S. E., & Lafferty, J. (2004). Mixed membership models of scientific publications. In *PANS-2004* (pp. 11885–11892).

Fujino, A., Ueda, N., & Nagata, M. (2010). A robust semi-supervised classification method for transfer learning. In *CIKM-2010* (pp. 379–388).

Gao, W., Cai, P., Wong, K. F., & Zhou, A. Y. (2010). Learning to rank only using training data from related domain. In *SIGIR-2010* (pp. 162–169).

Ghani, R. (2002). Combining labeled and unlabeled data for multi-class text categorization. In *ICML-2002* (pp. 187–194).

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *SIGIR-1999* (pp. 289–296).

Joachims, T. (1999). Transductive inference for text classification using support vector machines. In *ICML-1999* (pp. 200–209).

Joachims, T. (2003). Transductive learning via spectral graph partitioning. In *ICML-2003* (pp. 290–297).

John, S. T., & Nello, C. (2004). *Kernel methods for pattern analysis*. Cambridge University Press.

McCallum Nigam, K., Rennie, J., & Seymore, K. (2000). Automating the construction of internet portals with machine learning. *Information Retrieval Journal, 3*, 127–163.

Nigam, K., McCallum, A., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine Learning, 39*(2/3), 103–134.

Pan, W. K., Xiang, E. W., Liu Nathan, N., et al. (2010). Transfer learning in collaborative filtering for sparsity reduction. In *AAAI-2010* (pp. 230–235).

Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transaction on Knowledge and Data Engineering, 22*(10), 1345–1359.

Raina, R., Ng, A. Y., & Koller, D. (2006). Constructing informative priors using transfer learning. In *ICML-2006* (pp. 713–720).

Ruping, S., & Scheffer, T. (2005). Learning with multiple views. In *ICML-2005 workshop on learning with multiple views*.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management, 24*(5), 513–523.

Sarinnapakorn, K., & Kubat, M. (2007). Combining subclassifiers in text categorization: A dst-based solution and a case study. *IEEE Transaction Knowledge and Data Engineering, 19*(12), 1638–1651.

Tur, G. (2009). Co-adaptation: Adaptive co-training for semi-supervised learning. In *ICASSP-2009* (pp. 3721–3724).

Valiant, L. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.

Wang, P., Domeniconi, C., & Hu, J. (2008). Using wikipedia for co-clustering based cross-domain text classification. In: *ICDM-2008* (pp. 1085–1090).

Xiang, E. W., Cao, B., Hu, D. H., & Yang, Q. (2010). Bridging domains using world wide knowledge for transfer learning. *IEEE Transactions Knowledge Data Engineering (TKDE), 22*(6), 770–783.

Xue, G. R., Dai, W. Y., Yang, Q., & Yu, Y. (2008). Topic-bridged PLSA for cross-domain text classification. In *SIGIR-2008* (pp. 627–634).

Yang, Q., Chen, Y. Q., Xue, G. R., Dai, W. Y., & Yu, Y. (2009). Heterogeneous transfer learning for image clustering via the social Web. In *ACL/AFNLP2009* (pp. 1–9).

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *ACL-1995* (pp. 189–196).

Zhang, D., He, J. R., Liu, N., Si, L., & Lawrence, R. D. (2011). Multi-view transfer learning with a large margin approach. In *KDD-2011* (pp. 1208–1216).

Zhong, E. H., Fan, W., Peng, J., Zhang, K., Ren, J. T., Turaga, D. S., & Verscheure, O. (2009). Cross domain distribution adaptation via kernel mapping. In *KDD-2009* (pp. 1027–1036).

Zhu, S. H., Yu, K., Chi, Y., & Gong, Y. H. (2007). Combining content and link for classification using matrix factorization. In *SIGIR-2007* (pp. 487–494).

Zhu, X. J., & Goldberg, A. B. (2009). *Introduction to semi-supervised learning*. Morgan & Claypool Publishers.

Zhuang, F. Z., Luo, P., Shen, Z. Y., et al. (2010). Collaborative dual-PLSA: Mining distinction and commonality across multiple domains for Classification. In *CIKM-2010* (pp. 359–368).