Probabilistic Dyadic Data Analysis with Local and Global Consistency

Deng Cai[*] DENGCAI@CAD.ZJU.EDU.CN
Xuanhui Wang† XWANG20@CS.UIUC.EDU
Xiaofei He[*] XIAOFEIHE@CAD.ZJU.EDU.CN

[*]State Key Lab of CAD&CG, College of Computer Science, Zhejiang University, 100 Zijinggang Road, 310058, China
†Department of Computer Science, University of Illinois at Urbana-Champaign, 201 N. Goodwin Ave., Urbana, IL 61801.

Abstract

Dyadic data arises in many real world applications such as social network analysis and information retrieval. In order to discover the underlying or hidden structure in the dyadic data, many topic modeling techniques were proposed. The typical algorithms include Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). The probability density functions obtained by both of these two algorithms are supported on the Euclidean space. However, many previous studies have shown naturally occurring data may reside on or close to an underlying submanifold. We introduce a probabilistic framework for modeling both the topical and geometrical structure of the dyadic data that explicitly takes into account the local manifold structure. Specifically, the local manifold structure is modeled by a graph. The graph Laplacian, analogous to the Laplace-Beltrami operator on manifolds, is applied to smooth the probability density functions. As a result, the obtained probabilistic distributions are concentrated around the data manifold. Experimental results on real data sets demonstrate the effectiveness of the proposed approach.

1. Introduction

Dyadic data refers to domain where two sets of objects, row or column objects, are characterized by a matrix of numerical values which describe their mutual relationships. Such data arises in many real world applications such as social network analysis and information retrieval (Hofmann et al., 1998). A common example is term-document co-occurrence matrix. In order to discover the underlying or hidden structure in the dyadic data, topic modeling techniques are usually applied to learn a probabilistic interpretation of the row and column objects. Two of the most popular approaches for this purpose are Probabilistic Latent Semantic Indexing (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003).

In the dyadic aspect model applied to text analysis, a corpus of document is modeled as a set of pairs (d, w), where d is a document index and w is a word index. Each document is represented as a unique distribution over the k settings of the latent variable z. Each setting of the latent variable z corresponds to an underlying *topic*. Associated with each topic is a distribution over words in the vocabulary. Thus, a document is seen as a distribution over topics where each topic is described by a different distribution over words. A word is generated for a document by choosing a topic and then selecting a word according to the distribution over words for the chosen topic.

PLSA has been shown to be a low perplexity language model and outperforms Latent Semantic Indexing (LSI) (Deerwester et al., 1990) in terms of precision-recall on a number of document collections. However, the number of parameters of PLSA grows linearly with the number of documents, which suggests that PLSA is prone to overfitting (Blei et al., 2003). LDA was introduced to address this problem by incorporating a dirichlet regularization on the underlying topics. These two approaches do yield impressive results on exploratory dyadic data analysis. However, both of them fails take into account the geometry of the spaces where the objects (either column or row objects) reside. The learned probability distributions are simply supported on the ambient spaces.

Recent studies (Roweis & Saul, 2000; Belkin & Niyogi, 2001) have shown that naturally occurring data, such as texts and images, cannot possibly "fill up" the ambient Euclidean space, rather it must concentrate around lower dimensional structures. The goal of this paper is to extract

Appearing in *Proceedings of the 26^{th} International Conference on Machine Learning*, Montreal, Canada, 2009.

this kind of low dimensional structure and use it to regularize the learning of probability distributions. We construct a nearest neighbor graph to model the underlying manifold structure. The graph Laplacian, analogous to the Laplace-Beltrami operator on manifolds, is then used as a smoothing operator applied to the conditional probability distributions $p_z(z|d)$. This way, two sufficiently close documents should have similar conditional probability distributions. We use Kullback-Leibler divergence to measure the distance between two conditional probability distributions. The local consistency is incorporated into the probabilistic modeling framework through a regularizer. We discuss how to solve the regularized log-likelihood maximization problem using Expectation-Maximization techniques.

The rest of the paper is organized as follows. Section 2 provide a background of dyadic data analysis. Our Locally-consistent Topic Modeling (LTM) approach is introduced in Section 3. A variety experimental results are presented in Section 4. Finally, we give concluding remarks in Section 5.

2. Background

One of the popular approaches for dyadic data analysis is topic modeling. Recently, topic modeling algorithm receives a lot of interests (Li & McCallum, 2006; Rosen-Zvi et al., 2004). Two of the most well known topic modeling algorithms include Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 2001) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In the following of the paper, we will use text analysis to explain the algorithms.

The core of PLSA is a latent variable model for co-occurrence data which associates an unobserved topic variable $z_k \in \{z_1, \cdots, z_K\}$ with the occurrence of a word $w_j \in \{w_1, \cdots, w_M\}$ in a particular document $d_i \in \{d_1, \cdots, d_N\}$. As a generative model for word/document co-occurrences, PLSA is defined by the following scheme:

1. Select a document d_i with probability $P(d_i)$;
2. Pick a latent topic z_k with probability $P(z_k|d_i)$;
3. Generate a word w_j with probability $P(w_j|z_k)$.

As a result one obtains an observation pair (d_i, w_j), while the latent topic variable z_k is discarded. Translating the data generation process into a joint probability model results in the expression

$$\begin{aligned} P(d_i, w_j) &= P(d_i)P(w_j|d_i), \\ P(w_j|d_i) &= \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i). \end{aligned} \tag{1}$$

The parameters can be estimated by maximizing the log-likelihood

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log P(d_i, w_j) \\ &\propto \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \end{aligned} \tag{2}$$

where $n(d_i, w_j)$ the number of occurrences of term w_j in document d_i. The above optimization problem can be solved by using standard EM algorithm. Notice that there are $NK + MK$ parameters $\{P(w_j|z_k), P(z_k|d_i)\}$ which are independently estimated in PLSA model. It is easy to see that the number of parameters in PLSA grows linearly with the number of training documents (N). The linear growth in parameters suggests that the model is prone to overfitting (Blei et al., 2003).

To address this issue, Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is then proposed. LDA assumes that the probability distributions of documents over topics are generated from the same Dirichlet distribution with K parameters. The $K + MK$ parameters in a K-topic LDA model do not grow with the size of the corpus. Thus, LDA does not suffer from the same overfitting issue as PLSA.

3. Probabilistic Topic Modeling with Local Consistency

Recent studies (Roweis & Saul, 2000; Belkin & Niyogi, 2001) have shown that naturally occurring data, such as texts and images, cannot possibly "fill up" the ambient Euclidean space, rather it must concentrate around lower dimensional structures. In this section, we describe a principled way to extract this kind of low dimensional structure and use it to regularize the learning of probability distributions.

3.1. The Latent Variable Model with Graph Regularization

Recall that the documents $d \in D$ are drawn according to the distribution P_D. One might hope that knowledge of the distribution P_D can be exploited for better estimation of the conditional distribution $P(z|d)$. Nevertheless, if there is no identifiable relation between P_D and the conditional distribution $P(z|d)$, the knowledge of P_D is unlikely to be very useful.

Therefore, we will make a specific assumption about the connection between P_D and the conditional distribution $P(z|d)$. We assume that if two documents $d_1, d_2 \in D$ are *close* in the *intrinsic* geometry of P_D, then the conditional distributions $P(z|d_1)$ and $P(z|d_2)$ are "similar" to each other. In other words, the conditional probability

distribution $P(z|d)$ varies smoothly along the geodesics in the intrinsic geometry of P_D. This assumption is also referred to as *manifold assumption* (Belkin & Niyogi, 2001), which plays an essential rule in developing various kinds of algorithms including dimensionality reduction algorithms (Belkin & Niyogi, 2001) and semi-supervised learning algorithms (Belkin et al., 2006; Zhu & Lafferty, 2005).

Now we are facing two questions: 1.) how to measure the distance between two distributions? and 2.) how to model the local geometric structure in the data?

A popular way to measure the "distance" between two distributions is by using Kullback-Leibler Divergence (KL-Divergence). Given two distributions $P_i(z)$ and $P_s(z)$, the KL-Divergence between these two distributions is defined as:

$$D\Big(P_i(z)||P_s(z)\Big) = \sum_z P_i(z) \log \frac{P_i(z)}{P_s(z)} \tag{3}$$

It is important to note that KL-Divergence is not a distance measure because it is not symmetric.

Recent studies on spectral graph theory (Chung, 1997) and manifold learning theory (Belkin & Niyogi, 2001) have demonstrated that the local geometric structure can be effectively modeled through a nearest neighbor graph on a scatter of data points. Consider a graph with N vertices where each vertex corresponds to a document in the corpus. Define the edge weight matrix W as follows:

$$W_{is} = \begin{cases} 1, & \text{if } d_i \in N_p(d_s) \text{ or } d_s \in N_p(d_i) \\ 0, & \text{otherwise.} \end{cases} \tag{4}$$

where $N_p(d_i)$ denotes the set of p nearest neighbors of d_s (with respect to the Euclidean distance). It is important to note that W can be constructed to incorporate more information. For example, we can naturally use the label information of part of the data. If we know the labels of both d_i and d_s, we can set $W_{is} = 1$ when d_i and d_s share the same label and set $W_{is} = 0$ if d_i and d_s belong to different classes.

Let $P_i(z) \doteq P(z|d_i)$, the following term can be used to measure the smoothness of the conditional probability distribution $P(z|d)$ varies smoothly along the geodesics in the intrinsic geometry of data.

$$\mathcal{R} = \frac{1}{2} \sum_{i,s=1}^{N} \Big(D\big(P_i(z)||P_s(z)\big) + D\big(P_s(z)||P_i(z)\big)\Big) W_{is}. \tag{5}$$

By minimizing $\mathcal{R}$, we get a conditional probability distribution which is sufficiently smooth on the intrinsic document geometric structure. A intuitive explanation of minimizing $\mathcal{R}$ is that if two documents d_i and d_s are close (*i.e.* W_{is} is big), the distribution $P(z|d_i)$ and $P(z|d_s)$ are similar to each other.

Now we can define our new latent variable model. The new model adopts the generative scheme of PLSA. It aims to maximize the *regularized* log-likelihood as follows:

$$\begin{aligned} \mathfrak{L} &= \mathcal{L} - \lambda \mathcal{R} \\ &\propto \sum_{i=1}^{N} \sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k) P(z_k|d_i) \\ &\quad - \frac{\lambda}{2} \sum_{i,s=1}^{N} \Big(D\big(P_i(z)||P_s(z)\big) + D\big(P_s(z)||P_i(z)\big)\Big) W_{is} \end{aligned} \tag{6}$$

where λ is the regularization parameter.

Since this approach incorporates local consistency through a regularizer, we call it Locally-consistent Topic Modeling (LTM). It is important to note that this work is motivated from our previous work LapPLSA (Cai et al., 2008; Mei et al., 2008). The major difference is that LapPLSA constructs the regularizer using Euclidean distance. While in this work, we use the divergence measure which leads to a new objective function. We show how to apply EM algorithm to solve the optimization problem.

3.2. Model Fitting with EM

The standard procedure for maximum likelihood estimation in latent variable models is the Expectation Maximization (EM) algorithm (Dempster et al., 1977). EM alternates two steps: (i) an expectation (E) step where posterior probabilities are computed for the latent variables, based on the current estimates of the parameters, (ii) a maximization (M) step, where parameters are updated based on maximizing the so-called expected complete data log-likelihood which depends on the posterior probabilities computed in the E-step.

Same as PLSA, we also have $NK + MK$ parameters $\{P(w_j|z_k), P(z_k|d_i)\}$ and the latent variables are the hidden topics z_k in LTM. For simplicity, we use Ψ to denote all the $NK + MK$ parameters.

E-step:

The E-step for LTM is exactly same as the E-step in PLSA. The posterior probabilities for the latent variables are $P(z_k|d_i, w_j)$, which can be computed by simply applying Bayes' formula on Eq. (1)(Hofmann, 2001):

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{l=1}^{K} P(w_j|z_l)P(z_l|d_i)} \tag{7}$$

M-step:

With simple derivations (Hofmann, 2001), one can obtain the relevant part of the expected *complete* data log-

likelihood for LTM:

$$\begin{aligned} &Q(\Psi) = Q_1(\Psi) + Q_2(\Psi) \\ =&\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \sum_{k=1}^{K} P(z_k|d_i, w_j) \log\left[P(w_j|z_k)P(z_k|d_i)\right] \\ &- \frac{\lambda}{2}\sum_{i,s=1}^{N}\Big(D\big(P_i(z)||P_s(z)\big) + D\big(P_s(z)||P_i(z)\big)\Big)W_{is} \end{aligned} \tag{8}$$

To obtain the M-step re-estimation equations, we need to maximize $Q(\Psi)$ with respect to the parameters Ψ and with the constraints that $\sum_{k=1}^{K} P(z_k|d_i) = 1$ and $\sum_{j=1}^{M} P(w_j|z_k) = 1$.

Notice that $Q(\Psi)$ has two parts. The first part is exactly the expected complete data log-likelihood for PLSA. The second part is the regularization part which only involves the parameters $\{P(z_k|d_i)\}$. Thus, the M-step re-estimation equation for $\{P(w_j|z_k)\}$ will be exactly same as that in PLSA. It is(Hofmann, 2001):

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{m=1}^{M}\sum_{i=1}^{N} n(d_i, w_m)P(z_k|d_i, w_m)}, \tag{9}$$

Now let us derive the re-estimation equation for $\{P(z_k|d_i)\}$. In order to take care of the normalization constraints, Eq. (8) has to be augmented by appropriate Lagrange multipliers ρ_i,

$$\mathcal{H} = Q(\Psi) + \sum_{i=1}^{N}\rho_i\left(1 - \sum_{k=1}^{K} P(z_k|d_i)\right) \tag{10}$$

Maximization of $\mathcal{H}$ with respect to $\{P(z_k|d_i)\}$ leads to the following set of stationary equations

$$\begin{aligned} &\frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{P(z_k|d_i)} - \rho_i \\ &- \frac{\lambda}{2}\sum_{s=1}^{N}\left(\log\frac{P(z_k|d_i)}{P(z_k|d_s)} + 1 - \frac{P(z_k|d_s)}{P(z_k|d_i)}\right)W_{is} = 0, \\ &\qquad 1 \le i \le N, \quad 1 \le k \le K \end{aligned} \tag{11}$$

Because of the $\log$ term in the regularization part, it is hard to solve the above equations system. Recall the motivation of the regularization term, we hope that if two documents d_i and d_s are close (*i.e.* W_{is} is big), the distribution $P(z|d_i)$ and $P(z|d_s)$ are similar to each other, *i.e.*, $P(z_k|d_i)$ will be close to $P(z_k|d_s)$ and

$$\left(\frac{P(z_k|d_i)}{P(z_k|d_s)}\right)^{W_{is}} \approx 1.$$

Thus, we can use the following approximation:

$$\log(x) \approx 1 - \frac{1}{x}, \quad x \to 1.$$

The above approximation is based on the first order expansion of Taylor series of $\log$ function. With this approximation, the equations in Eq. (11) can be written as

$$\begin{aligned} &\frac{\sum_{j=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j)}{P(z_k|d_i)} - \rho_i \\ &- \frac{\lambda}{P(z_k|d_i)}\sum_{s=1}^{N}\Big(P(z_k|d_i) - P(z_k|d_s)\Big)W_{is} = 0, \\ &\qquad 1 \le i \le N, \quad 1 \le k \le K \end{aligned} \tag{12}$$

We have:

$$\begin{aligned} &\sum_{s=1}^{N}\Big(P(z_k|d_i) - P(z_k|d_s)\Big)W_{is} \\ =&P(z_k|d_i)\sum_{s=1}^{N} W_{is} - \sum_{s=1}^{N} P(z_k|d_s)W_{is} \end{aligned} \tag{13}$$

Let D denote a diagonal matrix whose entries are column (or row, since W is symmetric) sums of W, $D_{ii} = \sum_s W_{is}$. Define $L = D - W$, L is usually referred as graph Laplacian (Chung, 1997). We also define vector $\mathbf{y}_k = [P(z_k|d_1), \cdots, P(z_k|d_N)]^T$. It is easy to verify that Eq. (13) equals to the i-th element of vector $L\mathbf{y}_k$.

Let Ω denote a $n \times n$ diagonal matrix whose entries are ρ_i. The equations system in Eq. (12) can be rewritten as

$$\begin{aligned} &\begin{bmatrix} \sum_{j=1}^{M} n(d_1, w_j)P(z_k|d_1, w_j) \\ \vdots \\ \sum_{j=1}^{M} n(d_N, w_j)P(z_k|d_N, w_j) \end{bmatrix} - \Omega\mathbf{y}_k - \lambda L\mathbf{y}_k = 0, \\ &\qquad 1 \le k \le K. \end{aligned} \tag{14}$$

Let $\mathbf{e} \in \mathbb{R}^n$ denote the vector with all ones. With the normalization constraints, we know that $\sum_{k=1}^{K}\mathbf{y}_k = \mathbf{e}$. We can also easily verify that $L\mathbf{e} = \mathbf{0}$. Thus, add the K equations systems in Eq. (14) together, we can compute the Lagrange multipliers

$$\rho_i = \sum_{k=1}^{K}\sum_{j=1}^{M} n(d_i, w_j)P(z_k|d_i, w_j) = n(d_i),$$

where $n(d_i) = \sum_j n(d_i, w_j)$ refers to the document length.

One can easily verifies that the matrix $\Omega + \lambda L$ is positive definite. By solving the linear equations systems in

Eq. (14), one obtains the M-step re-estimation equation for $\{P(z_k|d_i)\}$

$$\mathbf{y}_k = (\Omega + \lambda L)^{-1} \begin{bmatrix} \sum_{j=1}^{M} n(d_1, w_j) P(z_k|d_1, w_j) \\ \vdots \\ \sum_{j=1}^{M} n(d_N, w_j) P(z_k|d_N, w_j) \end{bmatrix}. \quad (15)$$

When the regularization parameter $\lambda = 0$, we can easily see the above M-step re-estimation equation boils down to the M-step in original PLSA. $\Omega+\lambda L$ is usually a sparse matrix. Some efficient iterative algorithms (*e.g.*, LSQR (Paige & Saunders, 1982)) can be used to solve the above linear equations system instead of computing the matrix inversion.

The E-step (Eq. 7) and M-step (Eq. 9 and 15) are alternated until a termination condition is met.

4. Experiments

We evaluate our LTM approach in two application domains: document clustering and classification.

4.1. Document Clustering

Clustering is one of the most crucial techniques to organize the data in an unsupervised manner. The hidden topics extracted by the topic modeling approaches can be regarded as clusters. The estimated conditional probability density function $P(z_k|d_i)$ can be used to infer the cluster label of each datum. In this experiment, we investigate the use of topic modeling approach for text clustering.

4.1.1. Data and Experimental Settings

Our empirical study was conducted based on a subset of the Reuters-21578 text data set, provided by Reuters and corrected by Lewis.[1] 30 largest categories are chosen for our experiments, which includes 8,067 documents and 18,832 distinct words.

The clustering result is evaluated by comparing the obtained label of each document with that provided by the document corpus. The accuracy (AC) is used to measure the clustering performance (Xu et al., 2003). Given a document $\mathbf{x}_i$, let r_i and s_i be the obtained cluster label and the label provided by the corpus, respectively. The AC is defined as follows:

$$\text{AC} = \frac{\sum_{i=1}^{n} \delta(s_i, map(r_i))}{n}$$

where n is the total number of documents and $\delta(x, y)$ is the delta function that equals one if $x = y$ and equals zero otherwise, and map(r_i) is the permutation mapping function that maps each cluster label r_i to the equivalent label from the data corpus. The best mapping can be found by using the Kuhn-Munkres algorithm (Lovasz & Plummer, 1986).

In order to randomize the experiments, we conduct the evaluations with the cluster numbers ranging from two to ten. For each given cluster number k, 20 test runs were conducted on different randomly chosen clusters and we record both the average and standard deviation. We evaluate and compare three topic modeling algorithms and three traditional clustering algorithms as follows:

- Probabilistic latent semantic analysis (PLSA in short) (Hofmann, 2001).
- Latent dirichlet allocation (LDA in short) (Blei et al., 2003).
- Locally-consistent Topic Modeling (LTM in short). This is the method proposed in this paper.
- Kmeans clustering algorithm (Kmeans in short).
- Spectral clustering algorithm based on normalized cut criterion (NCut in short) (Shi & Malik, 2000; Ng et al., 2001).
- Nonnegative Matrix Factorization based clustering (NMF in short) (Xu et al., 2003).

There are two parameters in our LTM approach: the number of nearest neighbors p and the regularization parameter λ. Throughout our experiments, we empirically set the number of nearest neighbors p to 5, the value of the regularization parameter λ to 1000.

4.1.2. Performance Evaluation

Table 1 shows the clustering performance of the six approaches. We can see that both of the two traditional topic modeling approaches (PLSA and LDA) fail to achieve good performance (comparing to those standard clustering methods). One reason is that both PLSA and LDA discover the hidden topics in the Euclidean space and fail to consider the discriminant structure. By incorporating the geometric structure information into a graph regularizer and preserving the local consistency, the LTM approach gets significantly better performance than PLSA and LDA. The performance of LTM is also comparable with other three clustering algorithms. This shows that considering the intrinsic geometrical structure of the document space is important for learning a better hidden topic model in the sense of semantic structure.

Figure 1 shows how the performance of LTM varies with the parameters λ and p. The LTM is very stable with re-

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

Table 1. Clustering performance on Reuters

k	PLSA	LDA	LTM	kmeans	NCut	NMF
2	0.775±0.144	0.803±0.155	**0.892±0.128**	0.865±0.160	0.866±0.138	0.854±0.149
3	0.548±0.112	0.651±0.148	**0.814±0.108**	0.724±0.199	0.760±0.170	0.774±0.143
4	0.570±0.145	0.621±0.159	**0.775±0.133**	0.695±0.179	0.748±0.133	0.725±0.145
5	0.458±0.146	0.577±0.194	**0.712±0.141**	0.618±0.180	0.668±0.163	0.681±0.144
6	0.447±0.108	0.535±0.129	**0.675±0.115**	0.596±0.212	0.672±0.157	0.659±0.126
7	0.432±0.077	0.475±0.113	**0.644±0.113**	0.561±0.165	0.595±0.167	0.609±0.136
8	0.363±0.091	0.381±0.093	**0.534±0.097**	0.408±0.147	0.433±0.128	0.480±0.100
9	0.362±0.113	0.404±0.167	**0.596±0.101**	0.448±0.193	0.494±0.156	0.544±0.103
10	0.378±0.069	0.452±0.137	**0.576±0.092**	0.524±0.172	0.528±0.145	0.544±0.106
Avg.	0.481	0.544	**0.691**	0.604	0.641	0.652

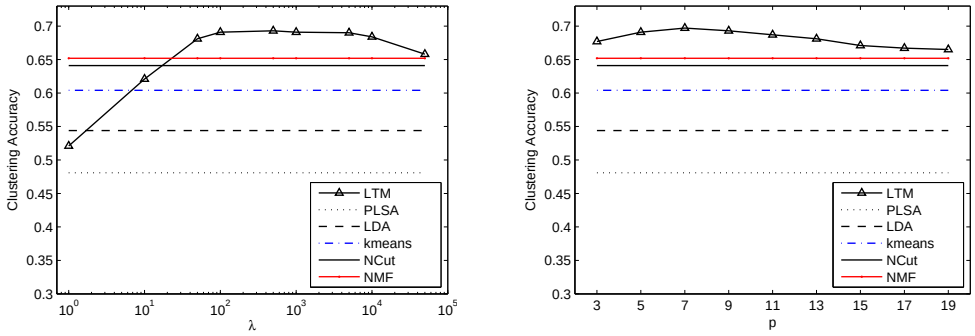

Figure 1. The performance of LTM vs. parameters λ and p. The LTM is very stable with respect to the parameter λ. It achieves consistent good performance with the λ varying from 100 to 10000. The performance is also stable when parameter p is between 3 than 9, then it decreases as the p increases.

spect to the parameter λ. It achieves consistent good performance with the λ varying from 100 to 10000. As we described, LTM uses a p-nearest graph to capture the local geometric structure of the data. It is more likely that a document share the same cluster membership with its p-nearest neighbor when p is small. Thus it is expected that the performance decreases as the p increases.

4.2. Document Classification

In the document classification problem, we wish to classify a document into two or more mutually exclusive categories. As in any classification problem, we may wish to consider generative approaches or discriminative approaches. In particular, by using one topic model for each class, we obtain a generative model for classification. It is also of interest to use topic modeling approaches (PLSA, LDA and LTM) in the discriminative framework, and this is our focus in this section.

A challenging aspect of the document classification problem is the choice of features. Treating individual words as features yields a rich but very large feature set (Joachims, 1998). One way to reduce this feature set is to use topic modeling approaches for dimensionality reduction. In particular, PLSA or LTM reduces any document to a fixed set of real-valued features $P(z_k|d_i)$. It is of interest to see how much discriminatory information we lose in reducing the document description to these parameters.

4.2.1. Data and Experimental Settings

The experimental settings in this work are basically the same as those in (Blei et al., 2003). Our empirical study was conducted based on a subset of the Nist Topic Detection and Tracking corpus (TDT2)[2]. This corpus consists of data collected during the first half of 1998 and taken from 6 sources, including 2 newswires (APW, NYT), 2 radio programs (VOA, PRI) and 2 television programs (CNN, ABC). We use the largest 10 categories in this experiment, which includes 7,456 documents and 33,947 distinct words.

In order to randomize the experiments, we conduct the evaluations with the training size for each category ranging from 2 to 20. For each case, 20 test runs were conducted on different randomly chosen labeled samples and we record both the average and standard deviation of the error rate.

[2]http://www.nist.gov/speech/tests/tdt/tdt98/index.htm

Table 2. Classification error rate on TDT2

	Word Feature	PLSA	LDA	LTM	LTM with Label
2	0.292±0.060	0.180±0.044	0.171±0.055	0.078±0.042	**0.074±0.040**
4	0.213±0.034	0.127±0.036	0.116±0.031	0.065±0.015	**0.057±0.010**
6	0.151±0.049	0.099±0.025	0.099±0.034	0.066±0.016	**0.057±0.008**
8	0.117±0.040	0.087±0.015	0.085±0.018	0.060±0.011	**0.057±0.009**
10	0.118±0.034	0.088±0.013	0.086±0.022	0.060±0.013	**0.055±0.007**
12	0.101±0.023	0.088±0.024	0.079±0.014	0.061±0.018	**0.056±0.010**
14	0.095±0.019	0.085±0.020	0.079±0.017	0.062±0.014	**0.059±0.011**
16	0.088±0.025	0.079±0.013	0.077±0.012	0.055±0.008	**0.054±0.006**
18	0.076±0.021	0.076±0.014	0.072±0.012	0.056±0.010	**0.053±0.007**
20	0.072±0.014	0.079±0.013	0.072±0.014	0.056±0.009	**0.054±0.007**
Avg.	0.132	0.099	0.094	0.062	**0.057**

In these experiments, we estimate the parameters of a PLSA (LDA) model on all the documents, without reference to their true class label. We then trained a support vector machine (SVM) on the low-dimensional representations and compared this SVM to an SVM trained on all the word features. For LTM, the label information of the training set can naturally be used to construct the graph. So we train two kinds of LTM models. One is purely un-supervised and we still use a p-nearest neighbor graph; the other utilizes the label information of the training set. We construct a graph in a semi-supervised manner, *i.e.*, we modify the p-nearest neighbor graph by removing edges between samples belonging to different categories and add edges between samples belonging to the same category.

4.2.2. Performance Evaluation

Table 2 shows the classification error rate of the five approaches. We can see that all the three topic modeling approaches gain improvement over the baseline (word feature), especially when the number of training sample is small. By incorporating the geometric structure information into a graph regularizer, the LTM approach gets significantly better performance than PLSA and LDA. When the label information (training set) is incorporated, LTM (with Label) obtains even better performance.

A key problem for all the topic modeling approaches is how to estimate the number of hidden topics. Figure 2 shows how the performance of three approaches varies with the number of the topics. Comparing to PLSA and LDA, the LTM model is less sensitive to the number of topics. This is another merit of applying LTM.

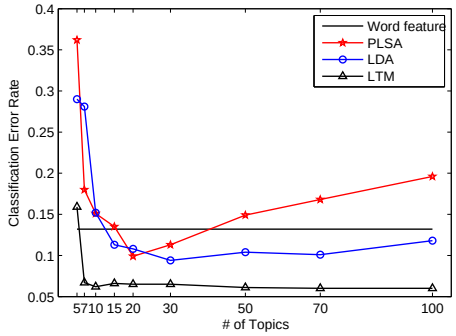

Figure 2. The performance of three algorithms vary with the number of topics.

5. Conclusion

We have presented a novel method for dyadic data analysis, called Locally-consistent Topic Modeling (LTM). LTM provides a principled way to incorporate the information in the intrinsic geometric structure of the data. Specifically, we adopt the popular manifold assumption and model the document space as a submanifold embedded in the ambient space and directly perform the topic modeling on this document manifold in question. As a result, LTM can have more discriminating power than traditional topic modeling approaches which discover the hidden topics in the Euclidean space, *e.g.* PLSA and LDA. Experimental results on text clustering and classification show that LTM provides better representation in the sense of semantic structure.

Several questions remain to be investigated in our future work:

1. There is a parameter λ which controls the smoothness of our LTM model. LTM boils down to original PLSA when $\lambda = 0$. Also, it is easy to see that $P(z_k|d_i)$ will be the same for all the documents when $\lambda = +\infty$. Thus, a suitable value of λ is critical to our algorithm. It remains unclear how to do model selection theoretically and efficiently.

2. We consider the topic modeling on document manifold and develop our approach based on PLSA. The idea of exploiting manifold structure can also be nat-

urally incorporated into other topic modeling algorithms, *e.g*., Latent Dirichlet Allocation.

3. It would be very interesting to explore different ways of constructing the document graph to incorporate other prior knowledge. There is no reason to believe that the nearest neighbor graph is the only or the most natural choice. For example, for web page data it may be more natural to use the hyperlink information to construct the graph.

Acknowledgments

This work was supported by the Program for Changjiang Scholars and Innovative Research Team in University (IRT0652, PCSIRT), National Science Foundation of China under Grant 60875044, National Key Basic Research Foundation of China under Grant 2009CB320801 and Yahoo! Ph.D Fellowship. Any opinions, findings, and conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Belkin, M., & Niyogi, P. (2001). Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in neural information processing systems 14*, 585–591. Cambridge, MA: MIT Press.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from examples. *Journal of Machine Learning Research*, *7*, 2399–2434.

Blei, D., Ng, A., & Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning Research*, 993–1022.

Cai, D., Mei, Q., Han, J., & Zhai, C. (2008). Modeling hidden topics on document manifold. *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management* (pp. 911–920).

Chung, F. R. K. (1997). *Spectral graph theory*, vol. 92 of *Regional Conference Series in Mathematics*. AMS.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W., & harshman, R. A. (1990). Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, *41*, 391–407.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proc. 1999 Int. Conf. on Research and Development in Information Retrieval* (pp. 50–57). Berkeley, CA.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, *42*, 177–196.

Hofmann, T., Puzicha, J., & Jordan, M. I. (1998). Learning from dyadic data. In *Advances in neural information processing systems 11*, 466–472.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *European conference on machine learning* (pp. 137–142).

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. *Proc. 2006 Int. Conf. Machine Learning* (pp. 577–584).

Lovasz, L., & Plummer, M. (1986). *Matching theory*. North Holland, Budapest: Akadémiai Kiadó.

Mei, Q., Cai, D., Zhang, D., & Zhai, C. (2008). Topic modeling with network regularization. *WWW '08: Proceeding of the 17th international conference on World Wide Web* (pp. 101–110).

Ng, A. Y., Jordan, M., & Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems 14*, 849–856. Cambridge, MA: MIT Press.

Paige, C. C., & Saunders, M. A. (1982). LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software*, *8*, 43–71.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence* (pp. 487–494).

Roweis, S., & Saul, L. (2000). Nonlinear dimensionality reduction by locally linear embedding. *Science*, *290*, 2323–2326.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*, 888–905.

Xu, W., Liu, X., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. *Proc. 2003 Int. Conf. on Research and Development in Information Retrieval (SIGIR'03)* (pp. 267–273). Toronto, Canada.

Zhu, X., & Lafferty, J. (2005). Harmonic mixtures: combining mixture models and graph-based methods for inductive and scalable semi-supervised learning. *ICML '05: Proceedings of the 22nd international conference on Machine learning* (pp. 1052–1059). Bonn, Germany.