Discriminative Topic Modeling based on Manifold Learning

Seungil Huh
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA
seungilh@cs.cmu.edu

Stephen E. Fienberg
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA
fienberg@stat.cmu.edu

ABSTRACT

Topic modeling has been popularly used for data analysis in various domains including text documents. Previous topic models, such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA), have shown impressive success in discovering low-rank hidden structures for modeling text documents. These models, however, do not take into account the manifold structure of data, which is generally informative for the non-linear dimensionality reduction mapping. More recent models, namely Laplacian PLSI (LapPLSI) and Locally-consistent Topic Model (LTM), have incorporated the local manifold structure into topic models and have shown the resulting benefits. But these approaches fall short of the full discriminating power of manifold learning as they only enhance the proximity between the low-rank representations of neighboring pairs without any consideration for non-neighboring pairs. In this paper, we propose Discriminative Topic Model (DTM) that separates non-neighboring pairs from each other in addition to bringing neighboring pairs closer together, thereby preserving the global manifold structure as well as improving the local consistency. We also present a novel model fitting algorithm based on the generalized EM and the concept of Pareto improvement. As a result, DTM achieves higher classification performance in a semi-supervised setting by effectively exposing the manifold structure of data. We provide empirical evidence on text corpora to demonstrate the success of DTM in terms of classification accuracy and robustness to parameters compared to state-of-the-art techniques.

Categories and Subject Descriptors

H.2.8 [**Database Application**]: Data mining; I.7.0 [**Document and Text Processing**]: General

General Terms

Algorithms

Keywords

Topic modeling, Dimensionality reduction, Document classification, Semi-supervised learning

KDD'10, July 25–28, 2010, Washington, DC, USA.

1. INTRODUCTION

Topic models are based on the notion that each data component (e.g., a document) can be represented by a mixture of basic components (or *topics*). In text analysis, topic models typically adopt the *bag-of-words* assumption that ignores the information from the ordering of words. Each document in a given corpus is thus represented by a histogram containing the occurrence of words. The histogram is modeled by a distribution over a certain number of topics, each of which is a distribution over words in the vocabulary. By learning the distributions, a corresponding low-rank representation of the high-dimensional histogram can be obtained for each document. Topic models, such as probabilistic Latent Semantic Analysis (pLSA) [11] and Latent Dirichlet Allocation (LDA) [4] have shown impressive empirical success by improving classification accuracy through the discovery of low-rank hidden structures. In addition, these models provide probabilistic interpretations of the generative process of data.

According to recent research [17, 14, 2], data from texts or images are often found to be placed on a low-rank non-linear manifold within the high-dimensional space of the original data. Therefore, learning the manifold structure can provide better dimensionality reduction mapping and visualization. Based on this assumption, several topic models were recently developed, namely, Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [6] and Locally-consistent Topic Modeling (LTM) [7]. Both of the topic models increase the proximity between the probability distributions of the data pairs with *favorable relationships* (i.e., within-class pairs or neighbors in manifolds) by adding the proximity as a regularization term to the log-likelihood function of pLSA. As a result, these models obtain probabilistic distributions concentrated around the manifold and show higher accuracy than pLSA and LDA for text clustering and classification tasks. However, LapPLSI and LTM fall short of the full discriminating power of manifold learning because the global manifold structure is often not well preserved only by enhancing the proximity between favorable pairs. The *unfavorable relationships* (i.e., between-class pairs or non-neighbors in manifolds) between data pairs should also be considered.

In this work, we propose a new topic model to focus more on discriminating power, which we refer to as Discriminative Topic Model (DTM). In order to address real-world problems in a semi-supervised setting (i.e., using a small amount of labeled data with a large amount of unlabeled data), DTM maintains the local consistency by considering the manifold structure of data as do LapPLSI and LTM. However, DTM

explicitly aims not only to increase the proximity between the probability distributions of the data pairs with favorable relationships, but also to increase the separability between those of the data pairs with unfavorable relationships. Due to the effectiveness of this more refined manifold learning formulation, DTM also preserves the global manifold structure, showing better performance in real-word document classification tasks than the previous approaches. We also present an efficient algorithm to solve the proposed regularized log-likelihood maximization problem based on the generalized Expectation-Maximization algorithm [10] and the concept of Pareto improvement [1]. Our model fitting algorithm does not require the regularization parameter to which the classification performance can be sensitive. We offer empirical evidence on two real world text corpora (20 newsgroups and Yahoo! News K-series) and demonstrate the superiority of DTM to state-of-the-art techniques.

The remainder of this paper is organized as follows. Section 2 provides the background and Section 3 overviews previous works. We then formulate DTM and describe how to fit the proposed model in Section 4. The experimental results with discussions are presented in Section 5, followed by conclusions in Section 6.

2. BACKGROUND AND NOTATIONS

We begin by describing the two basic components of our method: probabilistic Latent Semantic Analysis (pLSA) [11] as a topic model and Laplacian Eigenmaps [2] as a manifold learning algorithm.

2.1 Probabilistic Latent Semantic Analysis

One of the most well-known and fundamental topic models is probabilistic Latent Semantic Analysis (pLSA) [11]. Evolved from Latent Semantic Indexing (LSA) [9], pLSA defines a proper generative model based on a solid statistical foundation.

Suppose that we have a corpus that consists of N documents $\{d_1, d_2, \cdots, d_N\}$ with words from a vocabulary containing M words $\{w_1, w_2, \cdots, w_M\}$. In pLSA, the occurrence of a word w in a particular document d is associated with one of K unobserved topic variables $\{z_1, z_2, \cdots, z_K\}$. More formally, pLSA can be defined by the following generative process:

- select a document d with probability $P(d)$
- pick a latent class z with probability $P(z|d)$
- generate a word w with probability $P(w|z)$

By summing out the latent variable z, the joint probability of an observed pair (d, w) can be computed as

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

Based on this joint probability, we can calculate the log-likelihood as

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

where $n(d,w)$ denotes the number of times word w occurred in document d. Following the likelihood principle, one can determine $P(w|z)$ and $P(z|d)$ by maximizing the relevant part of Eq. (2):

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\log\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \tag{3}$$

2.2 Laplacian Eigenmaps

Traditional manifold learning algorithms [17, 14, 2] have given way to recent graph-based semi-supervised learning algorithms [19, 18, 3]. The goal of manifold learning is to recover the structure of a given dataset by non-linear mapping into a low-dimensional space. As a manifold learning algorithm, Laplacian Eigenmaps [2] was developed based on spectral graph theory [8].

Suppose that we have N data points $\{u_1, u_2, \cdots, u_N\}$, each of which is an $M \times 1$ vector. From the nearest neighbor graph of these data points, we define a *local similarity matrix* W that contains favorable pair-wise relationships among them:

$$W_{ij} = \begin{cases} 1, & \text{if } u_i \in \mathcal{N}_r(u_j) \text{ or } u_j \in \mathcal{N}_r(u_i) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

where $\mathcal{N}_r(u)$ is the set of the r nearest neighbors of u.

Let x_i, which is a $K \times 1$ vector, be a low-rank representation of u_i on the manifold (i.e., $K \ll N$). Intuitively, if two data points u_i and u_j are close to each other in the original space, the corresponding low-rank representations x_i and x_j should also lie near each other. From this intuition, Laplacian Eigenmaps minimize the following objective function:

$$\sum_{i,j=1}^{N} W_{ij}||x_i - x_j||^2 \tag{5}$$

This optimization problem, however, produces trivial solutions $x_1 = x_2 = \cdots = x_N$. To avoid these outcomes, we also need to somehow maintain unfavorable relationships between data points. For example, the original Laplacian Eigenmaps [2] impose the constraint, $XDX^T = I$ where X is the matrix, the i-th column of which is x_i and D is a diagonal matrix such that $D_{ii} = \sum_{j=1}^{n} W_{ij}$

3. PREVIOUS WORKS

Cai *et al.* recently proposed two topic models, Laplacian pLSI (LapPLSI) [6] and Locally-consistent Topic Modeling (LTM) [7], which use manifold structure information based on pLSA. To formalize these models, the objective of the Laplacian Eigenmaps is added as a regularization term to the original log-likelihood of pLSA. In LapPLSI, the Euclidean distance is adopted to measure the proximity between two probability distributions:

$$\begin{aligned} \mathcal{L} = &\sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i,w_j)\log\sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \\ &- \frac{\lambda}{2}\sum_{k=1}^{K}\sum_{i,j=1}^{N} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2 \end{aligned} \tag{6}$$

where λ is the regularization parameter and W is an $N \times N$ matrix measuring the local similarity of document pairs based on word occurrences.

In LTM, Kullback-Leibler Divergence is used instead of the Euclidean distance:

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i) \quad (7)$$
$$- \frac{\lambda}{2}\sum_{i,j=1}^{N} W_{ij}\Big(D\big(P(z|d_i)||P(z|d_j)\big) + D\big(P(z|d_j)||P(z|d_i)\big)\Big)$$

By discovering the local neighborhood structure, these two models show more discriminating power than pLSA and LDA for document clustering and classification tasks.

However, both of the models fall short of the full discriminating power of Laplacian Eigenmaps because the global manifold structure is often not well preserved only by enhancing the proximity between favorable pairs without maintaining or increasing the separability between unfavorable pairs. In addition, these models are limited in that their performance depends on the regularization parameter λ; futhermore, it is unclear how to appropriately determine its value.

4. DISCRIMINATIVE TOPIC MODEL

In this section, we formalize our proposed model, named Discriminative Topic Model (DTM). We also present an algorithm to solve the proposed regularized log-likelihood maximization problem based on the generalized Expectation Maximization (EM) algorithm [10] and the concept of Pareto improvement [1].

4.1 Regularized Model

When increasing the local consistency in manifold learning of data, we also need to maintain or increase the separability of the low-rank probability distributions of documents whose word occurrences are not close to each other. More formally, we need to minimize the proximity of the probability distributions of favorable pairs, expressed by

$$\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2 \quad (8)$$

Simultaneously, we need to maintain or maximize the separability of the probability distributions of unfavorable pairs, which can be expressed by

$$\sum_{i,j=1}^{N}\sum_{k=1}^{K}(1 - W_{ij})\big(P(z_k|d_i) - P(z_k|d_j)\big)^2 \quad (9)$$

Putting these two objectives together, we maximize the following objective function:

$$\frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}(1 - W_{ij})\big(P(z_k|d_i) - P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2} \quad (10)$$

which is equivalent to

$$\frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}(1 - W_{ij})\big(P(z_k|d_i) - P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2} + 1$$
$$= \frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2} \quad (11)$$

Our regularized model is regularized with this term to learn the manifold structure, in addition to adopting the generative process of pLSA. Thus, the log-likelihood of our model is as follows:

$$\mathcal{L} = \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j) \log \sum_{k=1}^{K} P(w_j|z_k)P(z_k|d_i)$$
$$+ \lambda\frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2} \quad (12)$$

where λ is the regularization parameter. Although this parameter is presented, due to the nature of our model fitting algorithm, this parameter need not be considered as we will elaborate in the following section.

4.2 Model Fitting

When a probabilistic model involves unobserved latent variables, the EM algorithm is generally used for the maximum likelihood estimation of the model. Here we use the generalized EM algorithm which in the M-step finds parameters that "improve" the expected value of the log-likelihood function rather than "maximizing" it. For more details, see [10].

Let $\phi = \{P(w_j|z_k)\}$ and $\theta = \{P(z_k|d_i)\}$, which are parameters of DTM. Thus, $MK + KN$ parameters are needed to be estimated, which is the same as pLSA.

E-step: The E-step of DTM is exactly the same as that of pLSA [11]. By applying Bayes' formula, we compute posterior probabilities.

$$P(z_k|d_i, w_j) = \frac{P(w_j|z_k)P(z_k|d_i)}{\sum_{k'=1}^{K} P(w_j|z_{k'})P(z_{k'}|d_i)} \quad (13)$$

M-step: In the M-step of DTM, we improve the expected value of the log-likelihood function which is

$$\mathcal{Q}(\phi, \theta) = \mathcal{Q}_1(\phi, \theta) + \lambda\mathcal{Q}_2(\theta) \quad (14)$$
$$= \sum_{i=1}^{N}\sum_{j=1}^{M} n(d_i, w_j)\sum_{k=1}^{K} P(z_k|d_i, w_j)\log[P(w_j|z_k)P(z_k|d_i)]$$
$$+ \lambda\frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i) - P(z_k|d_j)\big)^2}$$

The M-step re-estimation equations for ϕ are exactly the same as those of pLSA because the regularization term of DTM does not include $P(w_j|z_k)$.

$$P(w_j|z_k) = \frac{\sum_{i=1}^{N} n(d_i, w_j)P(z_k|d_i, w_j)}{\sum_{j'=1}^{M}\sum_{i=1}^{N} n(d_i, w_{j'})P(z_k|d_i, w_{j'})} \quad (15)$$

Before describing the M-step re-estimation algorithm for θ, we introduce the concept of Pareto improvement [1], based on which we propose our algorithm.

Pareto improvement is defined as a change from one status to another that can improve at least one objective without worsening any other objectives. More formally, in our problem, an update $\theta^{(t)} \rightarrow \theta^{(t+1)}$ is a Pareto improvement if either of the following two conditions is satisfied.

1. $\mathcal{Q}_1(\phi, \theta^{(t+1)}) > \mathcal{Q}_1(\phi, \theta^{(t)})$ and $\mathcal{Q}_2(\theta^{(t+1)}) \geq \mathcal{Q}_2(\theta^{(t)})$
2. $\mathcal{Q}_1(\phi, \theta^{(t+1)}) \geq \mathcal{Q}_1(\phi, \theta^{(t)})$ and $\mathcal{Q}_2(\theta^{(t+1)}) > \mathcal{Q}_2(\theta^{(t)})$

Based on the concepts of generalized EM and Pareto improvement, we re-estimate θ by 1) increasing $\mathcal{Q}(\phi, \theta)$ rather than maximizing it and 2) increasing at least one of $\mathcal{Q}_1(\phi, \theta)$ and $\mathcal{Q}_2(\theta)$ without decreasing the other.

One advantage of Pareto improvement is that $\mathcal{Q}(\phi,\theta)$ is improved regardless of the regularization parameter λ whose value affects the performance of previous models, and, yet is hard to determine appropriately.

In order to present a re-estimating algorithm for θ to increase $\mathcal{Q}(\phi,\theta)$ based on Pareto improvement, we first propose re-estimating equations to increase each of $\mathcal{Q}_1(\phi,\theta)$ and $\mathcal{Q}_2(\theta)$.

THEOREM 1. *If $\theta^{(t+1)}$ is computed from $\theta^{(t)}$ by applying the following re-estimation equations*

$$P(z_k|d_i) = \frac{\sum_{j=1}^{M} n(d_i,w_j)P(z_k|d_i,w_j)}{\sum_{j=1}^{M} n(d_i,w_j)} \tag{16}$$

then $\mathcal{Q}_1(\phi,\theta)$ monotonically increases when θ moves from $\theta^{(t)}$ to $\theta^{(t+1)}$ along the line with fixed ϕ.

PROOF. $\mathcal{Q}_1(\phi,\theta)$ is the expected value of the log-likelihood function of pLSA and Eq. (16) is the re-estimation equations for $P(z_k|d_i)$ of pLSA; thus, $\theta^{(t+1)}$ maximizes $\mathcal{Q}_1(\phi,\theta)$ when ϕ is fixed. Since $\mathcal{Q}_1(\phi,\theta)$ is a concave function of θ and $\theta^{(t+1)}$ is the maximum solution of $\mathcal{Q}_1(\phi,\theta)$, $\mathcal{Q}_1(\phi,\theta)$ monotonically increases when θ moves from $\theta^{(t)}$ to $\theta^{(t+1)}$ along the line. □

THEOREM 2. *Let α be the estimated value of the regularization term under the current estimates of the parameters: i.e.,*

$$\alpha = \frac{\sum_{i,j=1}^{N}\sum_{k=1}^{K}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2}{\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}\big(P(z_k|d_i)-P(z_k|d_j)\big)^2} \tag{17}$$

And we define β for topic id p *and document id* i *as*

$$\beta = \min\left(\frac{NP(z_p|d_i)+\alpha\sum_{j=1}^{N} W_{ij}P(z_p|d_j)}{\sum_{j=1}^{N} P(z_p|d_j)+\alpha D_{ii}P(z_p|d_i)}, \frac{1}{P(z_p|d_i)}\right) \tag{18}$$

Then, $\mathcal{Q}_2(\theta)$ is nondecreasing by the following re-estimation equations for $[P(z_1|d_i),P(z_2|d_i),\cdots,P(z_K|d_i)]$:

$$P(z_k|d_i) = \begin{cases} \beta P(z_p|d_i), & \text{if } k=p \\ \frac{1-\beta P(z_p|d_i)}{1-P(z_p|d_i)}P(z_k|d_i), & \text{otherwise} \end{cases} \tag{19}$$

PROOF. See Appendix A. □

It is worth mentioning that the minimum operator is inserted in Eq. (18) to ensure that $[P(z_1|d_i),\cdots,P(z_K|d_i)]$ becomes a probability distribution after re-estimation. It can be easily verified that $\sum_{k=1}^{K} P(z_k|d_i)=1$ and $\forall k$, $P(z_k|d_i)\geq 0$ after the re-estimation.

The re-estimation equations in Theorem 2 can be simplified and parallelized by matrix computation. See Appendix B for more details.

Now we propose our re-estimating algorithm for θ. Let the current parameters be θ_0. In order to maximize the discriminating power, we first compute θ_1 by repeatedly applying Eq. (19) to θ_0 with all possible pairs of (topic id, document id). Theorem 2 guarantees that $\mathcal{Q}_2(\theta_1)\geq\mathcal{Q}_2(\theta_0)$. We then test whether $\mathcal{Q}_1(\phi,\theta_1)\geq\mathcal{Q}_1(\phi,\theta_0)$, and if it is true, re-estimating for θ is done by setting $\theta=\theta_1$.

If $\mathcal{Q}_1(\phi,\theta_1)<\mathcal{Q}_1(\phi,\theta_0)$, θ is re-estimated through the local search from θ_1 as follows. θ_2 is computed from θ_1 by applying the E-step in Eq. (13) and the pLSA M-step in Eq. (16). Theorem 1 ensures that $\mathcal{Q}_1(\phi,\theta)$ monotonically increases when θ moves from θ_1 to θ_2 along the line. Thus, θ_3 is initially set as θ_1 and iterate the following update until $\mathcal{Q}_1(\phi,\theta_3)\geq\mathcal{Q}_1(\phi,\theta_0)$

$$\theta_3 = \theta_3 + \gamma(\theta_2-\theta_1) \tag{20}$$

where γ is the step parameter such that $0<\gamma\ll 1$.

We then test whether $\mathcal{Q}_2(\theta_3)\geq\mathcal{Q}_2(\theta_0)$. If it is true, re-estimating for θ is done by setting $\theta=\theta_3$. Otherwise, we keep θ as θ_0 without updating in the M-step and continue to the next E-step. Our model fitting algorithm is summarized in Algorithm 1.

It is worth discussing the role of the step parameter γ in Eq. (20). In some sense, γ plays a role in controlling the balance of the log-likelihood term $\mathcal{Q}_1(\phi,\theta)$ and the regularization term $\mathcal{Q}_2(\theta)$; the balance control is originally the role of the regularization parameter λ in Eq. (14). If γ is small, θ tends to be relatively close to θ_1; thus, the gap between two $\mathcal{Q}_1(\phi,\theta)$ in the consecutive iterations tends to be small, which leads to relatively large $\mathcal{Q}_1(\phi,\theta)$ and small $\mathcal{Q}_2(\theta)$ in the end of the fitting. Similarly, if γ is large, we can expect relatively small $\mathcal{Q}_1(\phi,\theta)$ and large $\mathcal{Q}_2(\theta)$ in the end of the fitting. Though γ influences the final value of $\mathcal{Q}_1$ and $\mathcal{Q}_2$, classification performance is not sensitive to γ as we will empirically show in the following section.

Algorithm 1 Model fitting for DTM

Input: $n(d_i,w_j)$: word occurrences in each document, N: # of documents, M: size of vocabulary, K: # of topics, W: similarity matrix, γ: step size, MI: max # of iterations.
Output: $\phi=\{P(w_j|z_k)\}$ and $\theta=\{P(z_k|d_i)\}$.

```
Randomly initialize φ and θ.
t ← 0
while t < MI do
  E-STEP:
  Compute P(z_k|d_i, w_j) using φ and θ as in Eq. (13).
  M-STEP:
  Re-estimate φ as in Eq. (15).
  θ_1 ← θ
  for p = 1 to K do
    for i = 1 to N do
      update P(z_1|d_i), ··· , P(z_K|d_i) in θ_1 with topic id
      p and document id i as in Eq. (19).
    end for
  end for
  if Q_1(φ, θ_1) ≥ Q_1(φ, θ) then
    Re-estimate θ by θ ← θ_1
  else
    Compute P(z_k|d_i, w_j) using φ and θ_1 as in Eq. (13).
    Compute θ_2 from θ_1 as in Eq. (16).
    θ_3 ← θ_1, s ← 0
    repeat
      θ_3 ← θ_3 + γ(θ_2 − θ1), s ← s + γ
    until Q_1(φ, θ_3) ≥ Q_1(φ, θ) or s + γ > 1
    if Q_1(φ, θ_3) ≥ Q_1(φ, θ) and Q_2(θ_3) ≥ Q_2(θ) then
      Re-estimate θ by θ ← θ_3
    end if
  end if
  t ← t + 1
end while
```

5. EXPERIMENTS

In this section, we evaluate the proposed DTM on the two widely used text corpora, 20 newsgroups and Yahoo! News K-series, in document classification.

5.1 Datasets and Experimental setup

The 20 newsgroups corpus is a collection of approximately 20,000 newsgroup documents, partitioned almost evenly across 20 different newsgroups [12]. The preprocessed version was downloaded from R. F. Corrêa's webpage[1]; this version includes 8,156 distinct words and is divided into a training set and a test set. Among the documents in the training set, we randomly select 100 documents from each category for each test run so that 2,000 documents are used for classification. Yahoo! News K-series is a collection of 2340 news articles belonging to one of 20 different categories [5]. The preprocessed version including 8,104 distinct words was downloaded from D. L. Boley's webpage[2]. Among all documents, we select the documents belonging to the category "Entertainment" and its sub-categories, using 1389 documents with 15 categories for every test run; these categories have varying sizes ranging from 278 to 9.

We evaluate the performance of DTM and provide comparison against previous topic models (including LapPLSI and LTM) and other traditional dimension reduction algorithms:

- Probabilistic Latent Semantic Analysis (pLSA) [11]
- Latent Dirichlet Allocation (LDA) [4]
- Laplacian Probabilistic Latent Semantic Indexing (LapPLSI) [6]
- Locally-consistent Topic Modeling (LTM) [7]
- Principal Component Analysis (PCA)
- Non-negative Matrix Factorization (NMF) [13].

Additionally, the approach using raw word histograms without any dimension reduction is tested.

To address real-world problems in a semi-supervised setting, we randomly select a small number of documents (one of 1, 3, 5, and 10) from each category as labeled data; the rest are considered to be unlabeled data. For each approach, we explore several numbers of topics or dimensionalities of the embedding space. For classification, a linear-kernel Support Vector Machine (SVM) is trained on the low-dimensional representations (i.e., $P(z_k|d_i)$ for topic models). After 20 test runs, the average of the accuracies is reported.

5.2 Implementation Details

The tf-idf weight scheme [15] is first applied to the word occurrences. The histogram intersection is then computed to measure the similarity of two documents after L1-normalization for each document. More formally, the similarity of two documents d_i and d_j is calculated as

$$\sum_{k=1}^{M} \min\left(\frac{n(d_i,w_k)}{n(d_i)}, \frac{n(d_j,w_k)}{n(d_j)}\right) \tag{21}$$

where $n(d,w)$ is the number of occurrences of word w and $n(d)$ is the total number of words in document d; i.e., $n(d) = \sum_{k=1}^{M} n(d,w_k)$. We found that this histogram intersection is as effective as the Euclidean distance in discovering the nearest neighbors of each document, which are in the same category.

We additionally utilize class label information when constructing the similarity matrix W, as described in the previous work [7]. More specifically, after generating a r-nearest neighbor graph in an unsupervised manner, we add edges between documents belonging to the same category and remove edges between documents belonging to different categories.

In our experiments, we set the number of the nearest neighbors r as 10, and the step parameter γ as 0.1. Although we chose these parameters, the classification performance is not sensitive to these parameters as we will show later in this section.

For performance comparison, we implemented the other approaches as follows. For pLSA, the source codes were downloaded from Peter Gehler's code and dataset page[3]. For LDA, Matlab Topic Modeling Toolbox 1.3.2[4] was used. For LapPLSI and LTM, the source codes were downloaded from the author's webpage[5]. We directly implemented the other two methods: PCA and NMF. The regularization parameters of LapPLSI and LTM were tuned to produce the best performance among 1, 10, 100, and 1000. All the other parameter settings and implementation details were set to be identical to DTM.

5.3 Results and Discussions

Figures 1 and 2 demonstrate that DTM consistently outperforms all other approaches, including the most recently proposed LapPLSI and LTM, in terms of classification accuracy. From these results, we can conclude that DTM is more successful in exposing the manifold structure inherent in 20 newsgroups and Yahoo! News K-series corpora. LapPLSI and LTM do not show such capabilities because they are not effective in preserving the global manifold, which is expected to be found in both the corpora since they comprise groups of highly related categories.

Among previous approaches, LapPLSI and LTM show higher performance than pLSA and LDA, as expected. Although LapPLSI and LTM do not reach the full discriminating power of manifold learning, they can still find a low-rank nonlinear embedding space to which documents are mapped. On the other hand, pLSA and LDA, which do not adopt any regularization for manifold learning, cannot find such a nonlinear embedding space. The performance of pLSA decreases as the number of topics increases beyond a certain point; it is well known that pLSA is prone to overfitting due to the large number of parameters which grows proportionally with data size. PCA and NMF also demonstrate similar tendencies on both of the corpora.

Figure 3 shows that DTM is insensitive to the variation in the two parameters utilized by the model: the number of the nearest neighbors r and the step size γ. Additionally, in contrast to LapPLSI and LTM, the regularization parameter is not needed in DTM. Therefore, DTM is negligibly affected by parameter changes.

[1] http://sites.google.com/site/renatocorrea02/textcategorizationdatasets/
[2] http://www-users.cs.umn.edu/~boley/ftp/PDDPdata/
[3] http://www.kyb.mpg.de/bs/people/pgehler/code/index.html
[4] http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm
[5] http://www.zjucadcg.cn/dengcai/LapPLSA/index.html

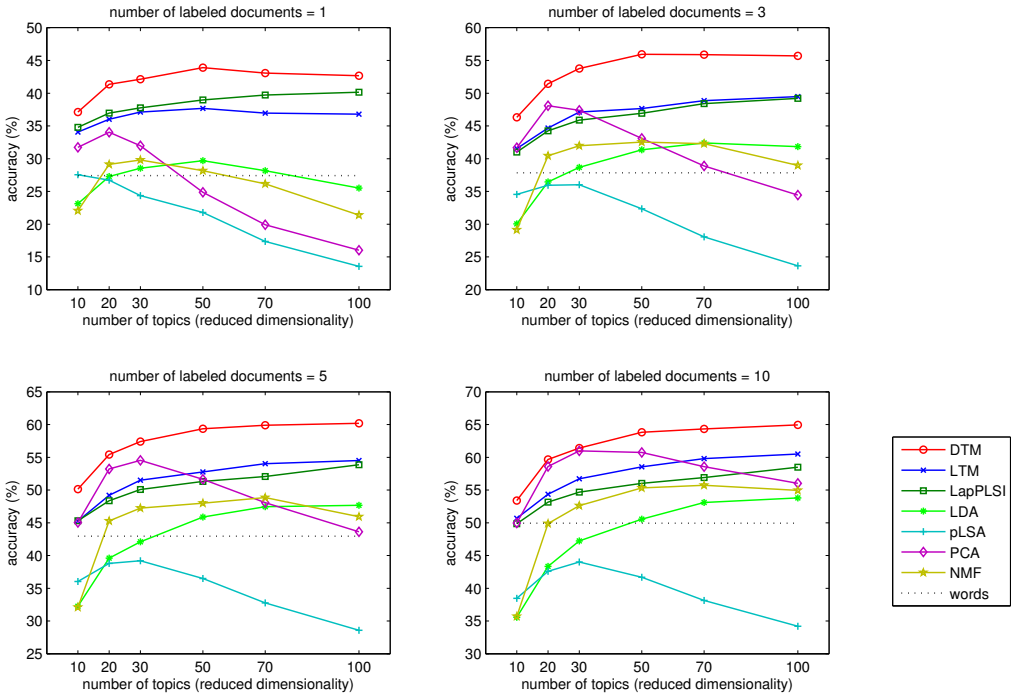

Figure 1: Classification performance on 20 newsgroups (best viewed in color)

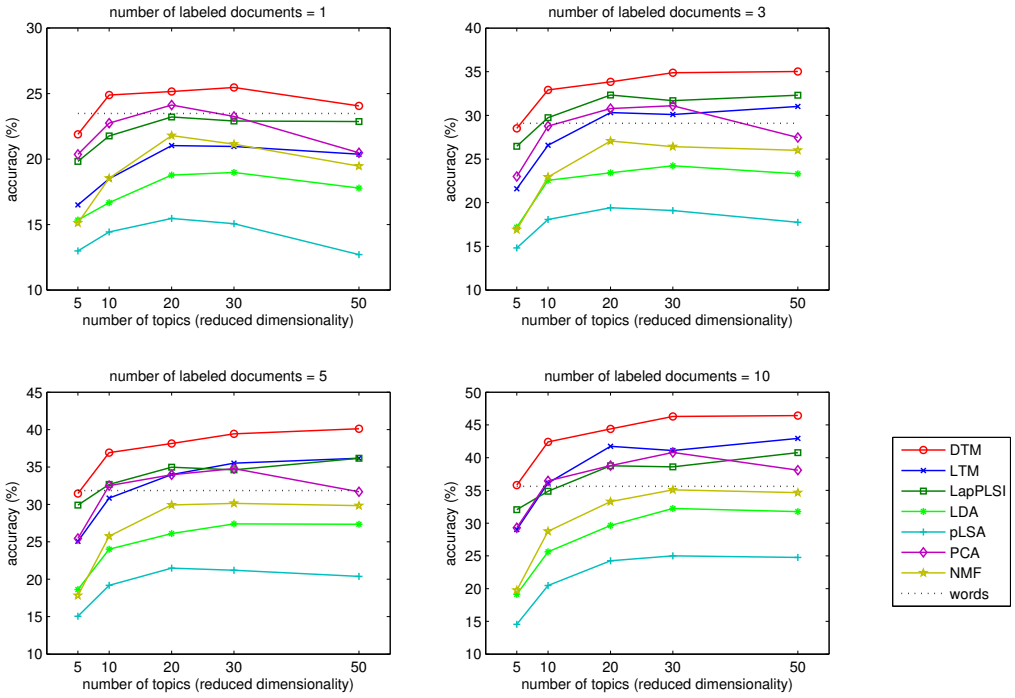

Figure 2: Classification performance on Yahoo! News K-series (best viewed in color)

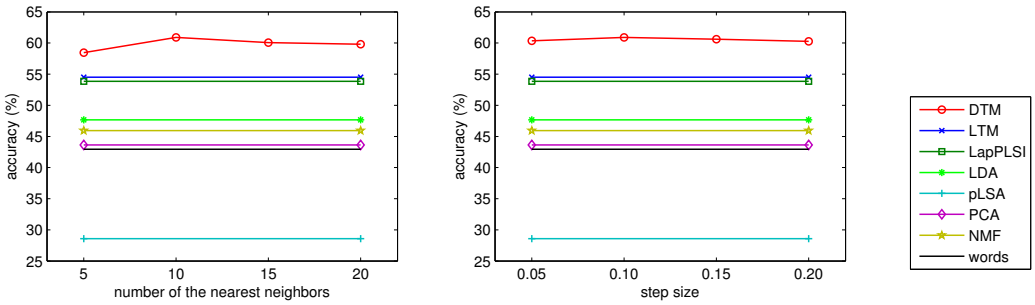

Figure 3: Classification performance of DTM as the parameters are varied (5 labeled data for each category and 100 topics on 20 newsgroups, best viewed in color)

6. CONCLUSIONS

In this paper, we have proposed a topic model that incorporates the information from manifold structures of data by considering unfavorable relationships in addition to favorable ones; the former are ignored in previous work. We have also presented an efficient model fitting algorithm, based on generalized EM and Pareto improvement, which enables reliable discovery of the low-rank hidden structures by minimizing the sensitivity to parameters. We empirically demonstrated that our approach ourperforms previous topic models in terms of classification accuracy in a semi-supervised setting on two text corpora.

7. ACKNOWLEDGMENTS

Seungil Huh is partially supported by a Samsung Fellowship. We thank the anonymous reviewers for their insightful feedback and Yoongu Kim for providing helpful comments for the final manuscript.

8. REFERENCES

[1] N. Barr. *Economics of the welfare state*. Oxford University Press, New York, USA, 2004.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *Advances in Neural Information Processing Systems*, volume 14, pages 586–691, 2001.

[3] M. Belkin, P. Niyogi, and V. Sindhwani. Mainfold regularization: A geometric framework for learning from examples. *Journal of Machine Learning*, 7:2399–2434, 2006.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning*, 3:993–1022, 2003.

[5] D. L. Boley. Principal direction divisive partitioning. In *Data Mining and Knowledge Discovery*, volume 2, pages 325–344, 1998.

[6] D. Cai, Q. Mei, J. Han, and C. Zhai. Modeling hidden topics on document manifold. In *Proceedings of the ACM conference on Information and knowledge management*, pages 911–920, 2008.

[7] D. Cai, X. Wang, and X. He. Probabilistic dyadic data analysis with local and global consistency. In *Proceedings of the International Conference on Machine Learning*, pages 105–112, 2009.

[8] F. R. K. Chung. *Spectral graph theory*. American Mathematical Society, 1997.

[9] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41:391–407, 1990.

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

[11] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of International Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[12] Home page for 20 newsgroups data set. http://people.csail.mit.edu/jrennie/20Newsgroups/.

[13] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2000.

[14] S. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[15] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24:513–523, 1988.

[16] F. Sha, L. K. Saul, and D. D. Lee. Multiplicative updates for nonnegative quadratic programming in support vector machines. In *Advances in Neural Information Processing Systems*, pages 1041–1048, 2003.

[17] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[18] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *Advances in Neural Information Processing Systems*, volume 14, pages 321–328, 2003.

[19] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning*, pages 912–919, 2003.

APPENDIX

A. Proof of Theorem 2

We reintroduce the concept of auxiliary function [13, 16].

DEFINITION 1. *$G(x, x')$ is an auxiliary function for $F(x)$ if the two following conditions are satisfied.*

$$G(x, x') \leq F(x), \quad G(x, x) = F(x) \tag{22}$$

This definition is useful with the following Lemma.

LEMMA 1. *If $G(x, x')$ is an auxiliary function, then $F(x)$ is non-increasing under the update*

$$x^{t+1} = \arg\max_{x} G(x, x') \tag{23}$$

PROOF. $F(x^{t+1}) \geq G(x^{t+1}, x^t) \geq G(x^t, x^t) = F(x^t)$. □

We define $\hat{\tau}$ for topic id p and document id i as

$$\hat{\tau} = \frac{NP(z_p|d_i) + \alpha \sum_{j=1}^{N} W_{ij} P(z_p|d_j)}{\sum_{j=1}^{N} P(z_p|d_j) + \alpha D_{ii} P(z_p|d_i)} \tag{24}$$

and also define

$$\mathcal{R}(\theta) = \sum_{i,j=1}^{N} \sum_{k=1}^{K} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 - \alpha \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 \tag{25}$$

LEMMA 2. *$\mathcal{R}(\theta)$ is nondecreasing after re-estimation of $[P(z_1|d_i), P(z_2|d_i), \cdots, P(z_K|d_i)]$ by the following equations with $\tau = \hat{\tau}$.*

$$P(z_k|d_i) = \begin{cases} \tau P(z_p|d_i), & \text{if } k = p \\ \frac{1-\tau P(z_p|d_i)}{1-P(z_p|d_i)} P(z_k|d_i), & \text{otherwise} \end{cases} \tag{26}$$

PROOF. Let $F(\tau)$ be the value of $\mathcal{R}(\theta)$ at $\theta = \tilde{\theta}^{(t+1)}$ that is obtained by applying the update in Eq. (26) to the current parameters $\theta^{(t)} = \{P(z|d)\}$. Then,

$$\frac{\partial F(\tau)}{\partial \tau} = 2 \sum_{j=1}^{N} (1 - \alpha W_{ij}) \Bigg[\big(\tau P(z_p|d_i) - P(z_p|d_j)\big) P(z_p|d_i) - \sum_{k \neq p} \Big(\frac{1 - \tau P(z_p|d_i)}{1 - P(z_p|d_i)} P(z_k|d_i) - P(z_k|d_j) \Big) \frac{P(z_p|d_i)}{1 - P(z_p|d_i)} \Bigg] \tag{27}$$

Since $\sum_{k \neq p} P(z_k|d) = 1 - P(z_p|d)$,

$$\frac{\partial F(\tau)}{\partial \tau} = 2c\big(NP(z_p|d_i) - \alpha D_{ii} P(z_p|d_i)\big)\tau - 2c\Big(\sum_{j=1}^{N} P(z_p|d_j) - \alpha \sum_{j=1}^{N} W_{ij} P(z_p|d_j) \Big) \tag{28}$$

where $c = \big(P(z_p|d_i) + \frac{P(z_p|d_i)}{1-P(z_p|d_i)}\big)$ and $D_{ii} = \sum_{j=1}^{N} W_{ij}$.

In addition, the second order derivative of $F(\tau)$ is

$$\frac{\partial^2 F(\tau)}{\partial \tau^2} = 2c\big(NP(z_p|d_i) - \alpha D_{ii} P(z_p|d_i)\big) \tag{29}$$

We define G as an auxiliary function of $F(\tau)$ by replacing the second order derivative in the Taylor series expansion of $F(\tau)$ at $\tau = 1$.

$$G(\tau, 1) = F(1) + \frac{\partial F(\tau)}{\partial \tau}\Big|_{\tau=1} (\tau - 1) - c\Big(\sum_{j=1}^{N} P(z_p|d_j) + \alpha D_{ii} P(z_p|d_i) \Big)(\tau - 1)^2 \tag{30}$$

Since $G(\tau, 1) - F(\tau) = -c\big(\sum_{j=1}^{N} P(z_p|d_j) + NP(z_p|d_i) \big)(\tau - 1)^2 \leq 0$, G is an auxiliary function of F. Solving $\frac{\partial G(\tau,1)}{\partial \tau} = 0$ yields $\hat{\tau}$ in Eq. (24), which minimizes $G(\tau, 1)$ because $G(\tau, 1)$ is concave with respect to τ. Therefore, by Lemma 1,

$$\mathcal{R}(\tilde{\theta}^{(t+1)}) = F(\hat{\tau}) \geq G(\hat{\tau}, 1) \geq G(1, 1) = F(1) = \mathcal{R}(\theta^{(t)}) \tag{31}$$

□

LEMMA 3. *$\mathcal{R}(\theta)$ is nondecreasing by the updates in Eq. (19) with β in Eq. (18).*

PROOF. For any μ such that $0 \leq \mu \leq 1$,

$$G(1, 1) = (1-\mu)G(1, 1) + \mu G(1, 1) \leq (1-\mu)G(1, 1) + \mu G(\hat{\tau}, 1) \tag{32}$$

Since $G(\tau, 1)$ is concave,

$$(1 - \mu)G(1, 1) + \mu G(\hat{\tau}, 1) \leq G((1 - \mu) + \mu\hat{\tau}, 1) \tag{33}$$

Thus, $G(1, 1) \leq G(\nu, 1)$ for any ν that is placed between 1 and $\hat{\tau}$ (either $1 \leq \nu \leq \hat{\tau}$ or $\hat{\tau} \leq \nu \leq 1$).

Let $\theta^{(t+1)}$ be obtained by applying the updates in Eq. (19) to $\theta^{(t)}$. Since β is always placed between 1 and $\hat{\tau}$,

$$\mathcal{R}(\theta^{(t+1)}) = F(\beta) \geq G(\beta, 1) \geq G(1, 1) = F(1) = \mathcal{R}(\theta^{(t)}) \tag{34}$$

□

Proof of Theorem 2

PROOF. Since $\alpha = \mathcal{Q}_2(\theta^{(t)})$, $\mathcal{R}(\theta^{(t)}) = 0$. By Lemma 3,

$$\mathcal{R}(\theta^{(t+1)}) = \sum_{i,j=1}^{N} \sum_{k=1}^{K} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 \Big|_{\theta=\theta^{(t+1)}} - \alpha \sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} \big(P(z_k|d_i)^{(t+1)} - P(z_k|d_j)\big)^2 \Big|_{\theta=\theta^{(t+1)}} \geq 0 \tag{35}$$

Therefore,

$$\mathcal{Q}_2(\theta^{(t+1)}) = \frac{\sum_{i,j=1}^{N} \sum_{k=1}^{K} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 \big|_{\theta=\theta^{(t+1)}}}{\sum_{i,j=1}^{N} \sum_{k=1}^{K} W_{ij} \big(P(z_k|d_i) - P(z_k|d_j)\big)^2 \big|_{\theta=\theta^{(t+1)}}} \geq \alpha = \mathcal{Q}_2(\theta^{(t)}) \tag{36}$$

□

B. Matrix Formulation of Re-estimation Equations in Theorem 2

Let P be a matrix such that $P_{ij} = P(z_i|d_j)$.

$$\sum_{i,j=1}^{N}\sum_{k=1}^{K} W_{ij}(P_{ki} - P_{kj})^2 = Tr(PLP^T) \tag{37}$$

where $L = D - W$, which is the graph Laplacian. In the same way,

$$\begin{aligned}\sum_{i,j=1}^{N}\sum_{k=1}^{K} \big(P_{ki} - P_{kj}\big)^2 &= Tr\big(P(NI_N - 1_N 1_N^T)P^T\big) \\ &= NTr(PP^T) - (P1_N)^T(P1_N)\end{aligned} \tag{38}$$

where I_N is the $N \times N$ identity matrix and 1_N is an N by 1 vector with all ones.

From Eqs. (37) and (38),

$$\alpha = \frac{NTr(PP^T) - (P1_N)^T(P1_N)}{Tr(PLP^T)} \tag{39}$$

Since L is a sparse matrix, α can be efficiently computed.

Now we reformalize β as a matrix form. For topic id p and document id i,

$$\beta_{pi} = \min\left(\frac{(NP + \alpha PW)_{pi}}{(P1_N 1_N^T + \alpha PD)_{pi}}, \frac{1}{P_{pi}}\right) \tag{40}$$

Considering all the documents with the topic id p, we define

$$\vec{\beta}_p = \min\left(\frac{(NP + \alpha PW)_p}{(P1_N 1_N^T + \alpha PD)_p}, \frac{1_N^T}{P_p}\right) \tag{41}$$

where X_p is the p-th row of matrix X and division is element-wise. Finally, for the topic id p, we obtain the following update for P.

$$P_k = \begin{cases} \vec{\beta}_p \otimes P_k, & \text{if } k = p \\ \frac{1_N^T - \vec{\beta}_p \otimes P_p}{1_N^T - P_p} \otimes P_k, & \text{otherwise} \end{cases} \tag{42}$$

where $\otimes$ denotes element-wise multiplication.