Probabilistic Latent Semantic Visualization: Topic Model for Visualizing Documents

Tomoharu Iwata   Takeshi Yamada   Naonori Ueda

NTT Communication Science Laboratories
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, Japan
{iwata,yamada,ueda}@cslab.kecl.ntt.co.jp

ABSTRACT

We propose a visualization method based on a topic model for discrete data such as documents. Unlike conventional visualization methods based on pairwise distances such as multi-dimensional scaling, we consider a mapping from the visualization space into the space of documents as a generative process of documents. In the model, both documents and topics are assumed to have latent coordinates in a two- or three-dimensional Euclidean space, or visualization space. The topic proportions of a document are determined by the distances between the document and the topics in the visualization space, and each word is drawn from one of the topics according to its topic proportions. A visualization, i.e. latent coordinates of documents, can be obtained by fitting the model to a given set of documents using the EM algorithm, resulting in documents with similar topics being embedded close together. We demonstrate the effectiveness of the proposed model by visualizing document and movie data sets, and quantitatively compare it with conventional visualization methods.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data Mining*; I.2.6 [**Artificial Intelligence**]: Learning; I.5.1 [**Pattern Recognition**]: Model—*Statistical*

General Terms

Algorithms

Keywords

Visualization, Topic model, Probabilistic latent semantic analysis

1. INTRODUCTION

Recently there has been great interest in topic models for analyzing documents and other discrete data. A topic model is a hierarchical probabilistic model, in which a document is modeled as a mixture of topics, where a topic is modeled as a probability distribution over words. Probabilistic Latent Semantic Analysis (PLSA) [10] and Latent Dirichlet Allocation (LDA) [4] are representative topic models, and they are used for a wide variety of applications such as information retrieval, text clustering and collaborative filtering.

In this paper, we propose a nonlinear visualization method based on a topic model, which we call *Probabilistic Latent Semantic Visualization* (PLSV). Visualization is a useful tool for understanding complex and high dimensional data, and it enables us to browse intuitively through huge amounts of data. A number of document visualization methods have been proposed [8, 22], and the importance of visualizing documents is increasing since documents such as web pages, blogs, e-mails, patents and scientific articles are being accumulated rapidly.

In PLSV, we consider a mapping from the visualization space into the space of documents as a generative process of documents. Both documents and topics are assumed to have latent coordinates in a two- or three-dimensional Euclidean space, or visualization space. The topic proportions of a document are determined by the Euclidean distances between the document coordinate and the topic coordinates. If a document is located near a topic, the probability that the document has the topic becomes high. Each word in a document is drawn from one of the topics according to its topic proportions, as in other topic models. The parameters in PLSV including latent coordinates of documents can be estimated by fitting the model to the given set of documents using the EM algorithm.

A number of visualization methods have been proposed, such as Multi-Dimensional Scaling (MDS) [20], Isomap [19] and Locally Linear Embedding (LLE) [17]. However, most of these methods take no account of the latent structure in the given data such as topics in the case of document data. For example, MDS embeds samples so that pairwise distances in the visualization space accurately reflect pairwise distances in the original space, and it does not consider topics explicitly. On the other hand, because PLSV considers topics, documents with similar semantics are embedded close together even if they do not share any words.

PLSA or LDA can extract a low-dimensional representation of a document as topic proportions. However, they are not appropriate for visualization since they cannot express

KDD'08, August 24–27, 2008, Las Vegas, Nevada, USA.

more than three topics in the two-dimensional visualization space, and the representation is an embedding in the simplex space but not in the Euclidean space. In contrast, PLSV can express any number of topics even with a two-dimensional representation, and it embeds documents in the Euclidean space, which is a metric space that most closely corresponds to our intuitive understanding of space. The topic proportions estimated by PLSA or LDA can be embedded in the Euclidean space by Parametric Embedding (PE) [11], which can employ a set of topic proportions as input. However, the topic proportions may not be suitable when they are embedded in the visualization space since these topics are estimated in a different space from the visualization space. Moreover, errors accumulated in the topic estimation process with PLSA or LDA cannot be corrected in the embedding process with PE since they are modularized, and this method may result in poor visualization. On the other hand, since PLSV simultaneously estimates topics and visualizes documents in one probabilistic framework, topics are estimated so as to be optimal when documents are embedded in the two- or three-dimensional visualization space.

The remainder of this paper is organized as follows. In Section 2, we formulate PLSV, and describe its parameter estimation procedures. In Section 3, we briefly review related work. In Section 4, we demonstrate the effectiveness of PLSV by visualizing document and movie data sets, and quantitatively compare it with conventional visualization methods. Finally, we present concluding remarks and a discussion of future work in Section 5.

2. PROBABILISTIC LATENT SEMANTIC VISUALIZATION

2.1 Model

In the following discussion, we assume that the given data are documents. However, PLSV is applicable to other discrete data, such as purchase logs in collaborative filtering and gene sequences in bioinformatics.

Suppose that we have a set of N documents $\boldsymbol{C} = \{\boldsymbol{w}_n\}_{n=1}^N$. Each document is represented by a sequence of M_n words denoted by $\boldsymbol{w}_n = (w_{n1}, \cdots, w_{nM_n})$, where $w_{nm} \in \{1, \cdots, W\}$ is the mth word in the sequence of the nth document, M_n is the number of words in the nth document, and W is the vocabulary size.

PLSV is a probabilistic topic model for finding the embedding of documents with coordinates $\boldsymbol{X} = \{\boldsymbol{x}_n\}_{n=1}^N$, where $\boldsymbol{x}_n = (x_{n1}, \cdots, x_{nD})$ is a coordinate of the nth document in the visualization space, and D is its dimensionality, usually $D = 2$ or 3. We assume that there are Z topics indexed by $\{z\}_{z=1}^Z$, and each topic has its associated coordinate $\boldsymbol{\phi}_z = (\phi_{z1}, \cdots, \phi_{zD})$ in the visualization space. The topic proportion of a document is determined by its Euclidean distances from topics in the visualization space as follows:

$$P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) = \frac{\exp\left(-\frac{1}{2} \parallel \boldsymbol{x}_n - \boldsymbol{\phi}_z \parallel^2\right)}{\sum_{z'=1}^Z \exp\left(-\frac{1}{2} \parallel \boldsymbol{x}_n - \boldsymbol{\phi}_{z'} \parallel^2\right)}, \quad (1)$$

where $P(z|\boldsymbol{x}_n, \boldsymbol{\Phi})$ is the zth topic proportion of the nth document, $\sum_{z=1}^Z P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) = 1$, $\boldsymbol{\Phi} = \{\boldsymbol{\phi}_z\}_{z=1}^Z$ is the set of topic coordinates, and $\parallel \cdot \parallel$ represents the Euclidean norm in the visualization space. If we assume this topic proportion, when the Euclidean distance between coordinates of document $\boldsymbol{x}_n$ and topic $\boldsymbol{\phi}_z$ is small, the topic proportion $P(z|\boldsymbol{x}_n, \boldsymbol{\Phi})$ becomes high. Since documents located close together in the visualization space are similar distances from topic coordinates, they have similar topic proportions.

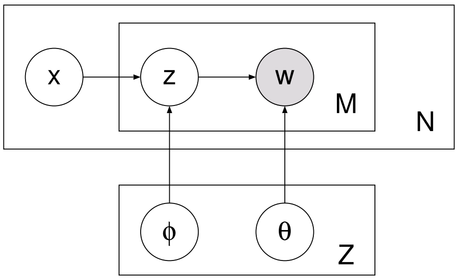


Figure 1: Graphical model representation of PLSV.

PLSV assumes the following generative process for a set of documents $\boldsymbol{C}$:

1. For each topic $z = 1, \cdots, Z$:
   (a) Draw word probability distribution
   $\boldsymbol{\theta}_z \sim \text{Dirichlet}(\alpha)$.
   (b) Draw topic coordinate
   $\boldsymbol{\phi}_z \sim \text{Normal}(\boldsymbol{0}, \beta^{-1}\boldsymbol{I})$.
2. For each document $n = 1, \cdots, N$:
   (a) Draw document coordinate
   $\boldsymbol{x}_n \sim \text{Normal}(\boldsymbol{0}, \gamma^{-1}\boldsymbol{I})$.
   (b) For each word $m = 1, \cdots, M_n$:
      i. Draw topic
      $z_{nm}|\boldsymbol{x}_n, \boldsymbol{\Phi} \sim \text{Mult}\Big(\{P(z|\boldsymbol{x}_n, \boldsymbol{\Phi})\}_{z=1}^Z\Big)$.
      ii. Draw word
      $w_{nm}|z_{nm}, \boldsymbol{\Theta} \sim \text{Mult}\Big(\{P(w|z_{nm}, \boldsymbol{\Theta})\}_{w=1}^W\Big)$.

Here $\boldsymbol{\Theta} = \{\boldsymbol{\theta}_z\}_{z=1}^Z$ is a set of word probabilities, $\boldsymbol{\theta}_z = \{\theta_{zw}\}_{w=1}^W$, $\sum_w \theta_{zw} = 1$, and $\theta_{zw} = P(w|z, \boldsymbol{\Theta})$ is the probability that the wth word occurs given the zth topic. As in LDA, each word w_{nm} is sampled from a topic-specific multinomial distribution, where the multinomial parameters $\boldsymbol{\theta}_z$ are generated by a Dirichlet distribution that is conjugate to multinomial. The coordinates $\boldsymbol{x}_n$ and $\boldsymbol{\phi}_z$ are assumed to be generated by zero-mean spherical Gaussian distributions for stabilizing the visualization. Given $\boldsymbol{x}_n$, $\boldsymbol{\Phi}$ and $\boldsymbol{\Theta}$, the probability of $\boldsymbol{w}_n$ is given as follows:

$$P(\boldsymbol{w}_n|\boldsymbol{x}_n, \boldsymbol{\Phi}, \boldsymbol{\Theta}) = \prod_{m=1}^{M_n} \sum_{z=1}^{Z} P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) P(w_{nm}|z, \boldsymbol{\Theta}). \quad (2)$$

Figure 1 shows a graphical model representation of PLSV, where shaded and unshaded nodes indicate observed and latent variables, respectively. Since each word in a document can be sampled from different topics, PLSV can capture multiple topics as in PLSA and LDA.

2.2 Parameter estimation

We estimate the parameters in PLSV based on maximum a posteriori (MAP) estimation. The unknown parameters are a set of document coordinates $\boldsymbol{X}$ and a set of topic coordinates $\boldsymbol{\Phi}$ as well as a set of word probabilities $\boldsymbol{\Theta}$. We represent all the unknown parameters by $\boldsymbol{\Psi} = \{\boldsymbol{X}, \boldsymbol{\Phi}, \boldsymbol{\Theta}\}$. The number of topics Z is assumed to be known and fixed.

The log likelihood of parameters $\boldsymbol{\Psi}$ given a set of documents $\boldsymbol{C}$ is as follows:

$$L(\boldsymbol{\Psi}|\boldsymbol{C}) = \sum_{n=1}^{N} \sum_{m=1}^{M_n} \log \sum_{z=1}^{Z} P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) P(w_{nm}|z, \boldsymbol{\Theta}). \quad (3)$$

Following the generative process described in the previous subsection, we use a Dirichlet prior for word probability $\boldsymbol{\theta}_z$:

$$p(\boldsymbol{\theta}_z) = \frac{\Gamma\big((\alpha+1)W\big)}{\Gamma(\alpha+1)^W} \prod_{w=1}^{W} \theta_{zw}^{\alpha}, \quad (4)$$

and a Gaussian prior with a zero mean and a spherical covariance for the coordinates of topic $\boldsymbol{\phi}_z$ and document $\boldsymbol{x}_n$:

$$p(\boldsymbol{\phi}_z) = \Big(\frac{\beta}{2\pi}\Big)^{\frac{D}{2}} \exp\Big(-\frac{\beta}{2} \parallel \boldsymbol{\phi}_z \parallel^2\Big), \quad (5)$$

$$p(\boldsymbol{x}_n) = \Big(\frac{\gamma}{2\pi}\Big)^{\frac{D}{2}} \exp\Big(-\frac{\gamma}{2} \parallel \boldsymbol{x}_n \parallel^2\Big), \quad (6)$$

where α, β and γ are hyper-parameters. We used $\alpha = 0.01$, $\beta = 0.1N$ and $\gamma = 0.1Z$ in all the experiments described in Section 4.

We estimate parameters $\boldsymbol{\Psi}$ by maximizing the posterior $p(\boldsymbol{\Psi}|\boldsymbol{C})$ using the EM algorithm [5]. The conditional expectation of the complete-data log likelihood with priors is represented as follows:

$$\begin{aligned} & Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}}) \\ = & \sum_{n=1}^{N} \sum_{m=1}^{M_n} \sum_{z=1}^{Z} P(z|n,m;\hat{\boldsymbol{\Psi}}) \log P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) P(w_{nm}|z, \boldsymbol{\Theta}) \\ + & \sum_{n=1}^{N} \log p(\boldsymbol{x}_n) + \sum_{z=1}^{Z} \log p(\boldsymbol{\phi}_z) + \sum_{z=1}^{Z} \log p(\boldsymbol{\theta}_z), \quad (7) \end{aligned}$$

where $\hat{\boldsymbol{\Psi}}$ represents the current estimate, and $P(z|n,m;\hat{\boldsymbol{\Psi}})$ represents the class posterior probability of the nth document and the mth word given the current estimate. In E-step, we compute the class posterior probability with the Bayes rule:

$$P(z|n,m;\hat{\boldsymbol{\Psi}}) = \frac{P(z|\hat{\boldsymbol{x}}_n, \hat{\boldsymbol{\Phi}}) P(w_{nm}|z, \hat{\boldsymbol{\Theta}})}{\sum_{z'=1}^{Z} P(z'|\hat{\boldsymbol{x}}_n, \hat{\boldsymbol{\Phi}}) P(w_{nm}|z', \hat{\boldsymbol{\Theta}})}, \quad (8)$$

where $P(z|\hat{\boldsymbol{x}}_n, \hat{\boldsymbol{\Phi}})$ is calculated by (1). In M-step, we obtain the next estimate of word probability $\hat{\theta}_{zw}$ by maximizing $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ w.r.t. θ_{zw} subject to $\sum_{w=1}^{W} \theta_{zw} = 1$:

$$\hat{\theta}_{zw} = \frac{\sum_{n=1}^{N} \sum_{m=1}^{M_n} I(w_{nm}=w) P(z|n,m;\hat{\boldsymbol{\Psi}}) + \alpha}{\sum_{w'=1}^{W} \sum_{n=1}^{N} \sum_{m=1}^{M_n} I(w_{nm}=w') P(z|n,m;\hat{\boldsymbol{\Psi}}) + \alpha W}, \quad (9)$$

where $I(\cdot)$ represents the indicator function, i.e. $I(A) = 1$ if A is true and 0 otherwise. The next estimates of document coordinate $\boldsymbol{x}_n$ and topic coordinate $\boldsymbol{\phi}_z$ cannot be solved in a closed form. Therefore, we estimate them by maximizing $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ using a gradient-based numerical optimization method such as the quasi-Newton method [15]. The gradients of $Q(\boldsymbol{\Psi}|\hat{\boldsymbol{\Psi}})$ w.r.t. $\boldsymbol{x}_n$ and $\boldsymbol{\phi}_z$ are respectively:

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{x}_n} = & \sum_{m=1}^{M_n} \sum_{z=1}^{Z} \Big(P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) - P(z|n,m;\hat{\boldsymbol{\Psi}})\Big)(\boldsymbol{x}_n - \boldsymbol{\phi}_z) \\ - & \gamma \boldsymbol{x}_n, \quad (10) \end{aligned}$$

$$\begin{aligned} \frac{\partial Q}{\partial \boldsymbol{\phi}_z} = & \sum_{n=1}^{N} \sum_{m=1}^{M_n} \Big(P(z|\boldsymbol{x}_n, \boldsymbol{\Phi}) - P(z|n,m;\hat{\boldsymbol{\Psi}})\Big)(\boldsymbol{\phi}_z - \boldsymbol{x}_n) \\ - & \beta \boldsymbol{\phi}_z. \quad (11) \end{aligned}$$

By iterating the E-step and the M-step until convergence, we obtain a local optimum solution for $\boldsymbol{\Psi}$. We can embed a new document with low computational cost by fixing topic coordinates $\hat{\boldsymbol{\Phi}}$ and word probabilities $\hat{\boldsymbol{\Theta}}$ to the estimated values.

3. RELATED WORK

3.1 PLSA

PLSV is based on Probabilistic Latent Semantic Analysis (PLSA) [10]. An essential difference between PLSV and PLSA is that a set of topic proportions are derived from coordinates of documents $\boldsymbol{X}$ and topics $\boldsymbol{\Phi}$ in PLSV, whereas they are directly estimated as $\boldsymbol{\Lambda} = \{\boldsymbol{\lambda}_n\}_{n=1}^{N}$ in PLSA, where $\boldsymbol{\lambda}_n = \{\lambda_{nz}\}_{z=1}^{Z}$, and $\lambda_{nz} = P(z|d_n, \boldsymbol{\Lambda})$ is the zth topic proportion of the nth document. Therefore although PLSV can express any number of topics in D-dimensional space, PLSA can express only $D+1$ topics in D-dimensional space. The number of parameters for topic proportions in PLSV is $(N+Z)D$, and it is much smaller than that in PLSA $N(Z-1)$ since usually $D \ll Z \ll N$. Therefore, PLSV can prevent overfitting compared with PLSA.

Under PLSA, the probability of $\boldsymbol{w}_n$ given d_n, $\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$ is as follows:

$$P(\boldsymbol{w}_n|d_n, \boldsymbol{\Lambda}, \boldsymbol{\Theta}) = \prod_{m=1}^{M_n} \sum_{z=1}^{Z} P(z|d_n, \boldsymbol{\Lambda}) P(w_{nm}|z, \boldsymbol{\Theta}). \quad (12)$$

The unknown parameters in PLSA $\boldsymbol{\Upsilon}$ are a set of topic proportions $\boldsymbol{\Lambda}$ and a set of word probabilities $\boldsymbol{\Theta}$. They can be estimated by maximizing the following likelihood with the EM algorithm:

$$L(\boldsymbol{\Upsilon}|\boldsymbol{C}) = \sum_{n=1}^{N} \sum_{m=1}^{M_n} \log \sum_{z=1}^{Z} P(z|d_n, \boldsymbol{\Lambda}) P(w_{nm}|z, \boldsymbol{\Theta}). \quad (13)$$

In E-step, the class posterior probability given the current estimate can be computed as follows:

$$P(z|d_n, w_{nm}; \hat{\boldsymbol{\Upsilon}}) = \frac{P(z|d_n, \hat{\boldsymbol{\Lambda}}) P(w_{nm}|z, \hat{\boldsymbol{\Theta}})}{\sum_{z'=1}^{Z} P(z'|d_n, \hat{\boldsymbol{\Lambda}}) P(w_{nm}|z', \hat{\boldsymbol{\Theta}})}. \quad (14)$$

In M-step, the next estimate of topic proportion $\hat{\lambda}_{nz}$ is given by:

$$\hat{\lambda}_{nz} = \frac{\sum_{m=1}^{M_n} P(z|d_n, w_{nm}; \hat{\boldsymbol{\Upsilon}})}{\sum_{z'=1}^{Z} \sum_{m=1}^{M_n} P(z'|d_n, w_{nm}; \hat{\boldsymbol{\Upsilon}})}, \quad (15)$$

and the next estimate of word probability $\hat{\theta}_{zw}$ is given by:

$$\hat{\theta}_{zw} = \frac{\sum_{n=1}^{N}\sum_{m=1}^{M_n} I(w_{nm}=w)P(z|d_n,w_{nw};\hat{\boldsymbol{\Upsilon}})}{\sum_{w'=1}^{W}\sum_{n=1}^{N}\sum_{m=1}^{M_n} I(w_{nm}=w')P(z|d_n,w_{nm};\hat{\boldsymbol{\Upsilon}})}. \tag{16}$$

We can use Dirichlet priors for the topic proportions and word probabilities as in PLSV.

3.2 Parametric Embedding

Parametric Embedding (PE) [11] is a nonlinear visualization method, which takes a set of discrete probability distributions as its input. The topic proportions estimated using PLSA $\hat{\boldsymbol{\Lambda}}$ with any number of topics can be embedded in a D-dimensional Euclidean space by PE. PE embeds samples in a low-dimensional Euclidean space so as to preserve the input probabilities by minimizing the following sum of Kullback-Leibler divergences:

$$E(\boldsymbol{X},\boldsymbol{\Phi}) = \sum_{n=1}^{N}\sum_{z=1}^{Z} P(z|d_n,\hat{\boldsymbol{\Lambda}})\log\frac{P(z|d_n,\hat{\boldsymbol{\Lambda}})}{P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})}, \tag{17}$$

where $P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})$ is the probability that the zth topic is chosen given a coordinate $\boldsymbol{x}_n$, and it is defined in the same equation as PLSV as in (1). The unknown parameters, a set of coordinates of documents $\boldsymbol{X}$ and topics $\boldsymbol{\Phi}$, can be obtained with a gradient-based numerical optimization method. The gradients of $E(\boldsymbol{X},\boldsymbol{\Phi})$ w.r.t. $\boldsymbol{x}_n$ and $\boldsymbol{\phi}_z$ are respectively:

$$\frac{\partial E}{\partial \boldsymbol{x}_n} = \sum_{z=1}^{Z}\Big(P(z|d_n,\hat{\boldsymbol{\Lambda}}) - P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})\Big)(\boldsymbol{x}_n - \boldsymbol{\phi}_z), \tag{18}$$

$$\frac{\partial E}{\partial \boldsymbol{\phi}_z} = \sum_{n=1}^{N}\Big(P(z|d_n,\hat{\boldsymbol{\Lambda}}) - P(z|\boldsymbol{x}_n,\boldsymbol{\Phi})\Big)(\boldsymbol{\phi}_z - \boldsymbol{x}_n). \tag{19}$$

We can use spherical Gaussian priors with zero means for the coordinates as in PLSV.

The parameter estimation procedures in PLSV described in Section 2.2 show that the calculation of coordinates in PE is included in the M-step in PLSA. These coordinates are then mapped to the topic proportions of documents and used in the E-step.

3.3 Other related work

There are a few visualization methods based on generative models. The examples include Generative Topographic Mapping (GTM) [1] and the visualization method proposed in [13]. However, these methods are not topic models, in which each word is assumed to be drawn from one of the topics according to the topic proportions. A topic model with latent coordinates based on GTM is proposed in [12]. However, it is mainly used for predicting sequences, and not for visualization. The correlated topic model is a topic model that can model correlations among topics [3]. PLSV can also model the topic correlations because topics embedded close together in the visualization space are likely to occur together. PLSV is an unsupervised visualization method, where topics are unobservable variables, and it is different from supervised visualization methods, such as Fisher linear discriminant analysis [7].

4. EXPERIMENTS

4.1 Compared methods

For evaluation, we compared PLSV with MDS, Isomap, PLSA and PLSA+PE by the visualization in the two-dimensional space $D = 2$.

MDS is a linear dimensionality reduction method, which embeds samples so as to minimize the discrepancy between pairwise distances in the visualization space and those in the original space. We used word count vectors as input $\boldsymbol{v}_n = (v_{n1},\cdots,v_{nW})$, where $v_{nw} = \sum_{m=1}^{M_n} I(w_{nm}=w)$ is the count of the wth word in the nth document. We normalized the vector so that the L-2 norm becomes one.

Isomap is a nonlinear dimensionality reduction method, in which a graph is first constructed by connecting h-nearest neighbors, and then samples are embedded so as to preserve the shortest path distances in the graph by MDS. We used the cosine similarity between word count vectors $\boldsymbol{v}_n$ for finding neighbors, which is widely used for the similarity measurement between documents. We set the number of neighbors at $h = 5$.

In PLSA, we visualized documents using the procedures described in Section 3.1. In order to represent topic proportions in a two-dimensional space, we converted the topic proportions estimated by PLSA with $Z = 3$ to a coordinate in the two-dimensional simplex.

PLSA+PE embeds documents using PE according to the topic proportions that are estimated by PLSA as described in Section 3.2.

4.2 Data sets

We used the following three data sets in the evaluations: NIPS, 20News and EachMovie.

NIPS data consist of papers from the NIPS conference from 2001 to 2003 [1]. There were 593 documents, and the vocabulary size was 14,036. Each document is labeled with 13 research areas, such as Neuroscience and Applications.

20News data consist of documents in the 20 Newsgroups corpus [14]. The corpus contains about 20,000 articles categorized into 20 discussion groups. We omitted stop-words and words that occurred fewer than 50 times, and also omitted documents with fewer than 50 words. The vocabulary size was 6,754. We sampled 50 documents from each of 20 classes, for a total of 1000 documents.

EachMovie data consist of movie ratings, which are standard benchmark data for collaborative filtering. We regarded movies and users as documents and words, respectively, where a movie is represented by a sequence of rated users. Each movie is labeled with 10 genres, for instance Action, Comedy and Romance. We omitted users and movies with fewer than 50 ratings and movies labeled with more than one genre. The number of movies was 764, and the number of users was 7,180.

4.3 Evaluation measurement

[1] Available at http://ai.stanford.edu/˜gal/

We evaluated the visualization results quantitatively from the label prediction accuracy with the k-nearest neighbor (k-NN) method in the visualization space. Each sample in all three data sets is labeled with research area, discussion group or genre. Note that we did not use the label information for visualization in any of the methods, namely, we performed visualization with fully unsupervised settings. The accuracy generally becomes high when samples with the same labels are located close together and samples with different labels are located far away from each other in the visualization space.

The k-NN method predicts a sample label from the most dominant label among the k nearest samples, where we used the Euclidean distance for finding neighbors. The accuracy is computed by $acc(k) = \frac{1}{N}\sum_{n=1}^{N} I\big(y_n = \hat{y}_k(\boldsymbol{x}_n)\big) \times 100$, where y_n is the label of the nth document, and $\hat{y}_k(\boldsymbol{x}_n)$ is the predicted label with the k-NN method for a sample with coordinate $\boldsymbol{x}_n$.

4.4 Results

The accuracies on NIPS, 20News and EachMovie data sets are shown in Figure 2 when documents are embedded in a two-dimensional space by PLSV, MDS, Isomap, PLSA and PLSA+PE. We set the number of topics at $Z = 50$ for PLSV and PLSA+PE. The number of topics for PLSA is automatically determined as $Z = 3$ since $D = 2$. In 20News, we created 30 evaluation sets by random sampling, where each set consists 1000 documents, and evaluated by the average accuracy over the 30 sets. In NIPS and EachMovie, the accuracies of PLSV, PLSA+PE and PLSA are averaged over 30 visualizations for one data set with different initial parameters. Only the standard deviations for PLSV are shown. In all three data sets, the highest accuracies are achieved by PLSV. This result implies that PLSV can appropriately embed documents in the two-dimensional Euclidean space while keeping the essential characteristics of the documents. The accuracies achieved by PLSA+PE are lower than those achieved by PLSV since it embeds documents through two modularized processes, where the objective functions are different from each other. The accuracies achieved by PLSA are low since it has only three topics, and it may be inadequate to measure similarities among topic proportions based on the Euclidean distance.

We also evaluated PLSV and PLSA+PE with different numbers of topics $Z = 5, 10, \cdots, 50$. The accuracies with the one-nearest neighbor method are shown in Figure 3. With a small number of topics, the accuracies achieved by PLSV are not very high, and they are comparable to those achieved by PLSA+PE. However, as the number of topics increases, PLSV outperforms PLSA+PE. This result indicates that the topic proportions and the word probabilities overfit the high dimensional PLSA parameter space in PLSA+PE, and they may not be appropriately represented in the two-dimensional visualization space.

Figures 4, 5 and 6 show visualization results obtained by PLSV ($Z = 50$), MDS, Isomap, PLSA ($Z = 3$) and PLSA+PE ($Z = 50$) on NIPS, 20News and EachMovie data sets, respectively. Here each point represents a document, and the shape and color represents the label. In the PLSV visualizations, documents with the same label are likely to be clustered together. On the other hand, with MDS and Isomap, documents with different labels are mixed, and thus the accuracy of their visualization is low. In PLSA, many documents are located at the corner, and the latent topic structure of the given data is not fully expressed in this dimensionality. In PLSA+PE, documents are slightly more mixed than those in PLSV as shown quantitatively by the accuracy.

Figure 7 shows PLSV visualization results for NIPS data with different numbers of topics $Z = 10, 20, 30$ and 40. Although documents with different labels are mixed when the number of topics is small, the visualization with $Z = 40$ shows similar quality to that with $Z = 50$.

We analyzed the PLSV visualization in detail. Figure 8 shows the visualization result for 20News data obtained by PLSV with $Z = 50$. Here each black circle represents a topic coordinate $\boldsymbol{\phi}_z$, and each black × represents a mean coordinate of documents for each label, which is calculated by $\boldsymbol{\mu}_y = \frac{1}{N_y}\sum_{n=1}^{N} I(y_n = y)\boldsymbol{x}_n$, where N_y is the number of documents labeled with y. The number near the circle corresponds to the topic index in the table at the bottom, where the ten most probable words for each topic are shown. Documents with the same label are clustered together, and closely related labels are located nearby, such as rec.sport.baseball and rec.sport.hockey, rec.autos and rec.motorcycles, comp.-sys.mac.hardware and comp.sys.ibm.pc.hardware, and soc.-religion.christian and talk.religion.misc. In a topic located near a label mean, representative words for the label occur with high probability. For example, probable words in topics near rec.sport ($z = 1$ and $z = 2$) are 'team', 'players' and 'game', those near com.graphics ($z = 5$) are 'graphics', 'points' and 'lines', and those near sci.crypt ($z = 8$) are 'clipper', 'encryption' and 'public'.

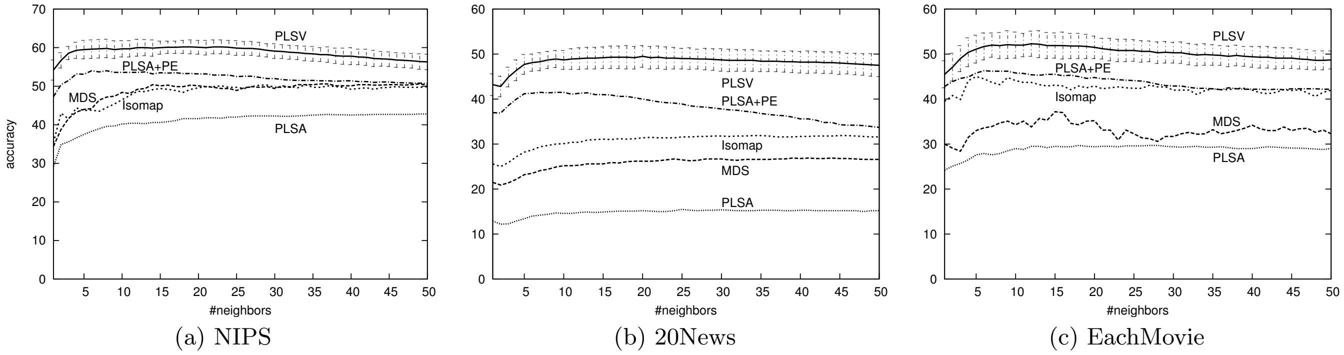
Figure 9 shows the visualization result for certain movie titles from EachMovie data obtained by PLSV with $Z = 50$. Movies in the same genre are likely to be located close together. For example, classic movies are located in the bottom right, and foreign movies are located at the top. Classic movies are tightly clustered because there may be a number of people who see only classic movies.

The computational time of PLSV on a PC with 3.2GHz Xeon CPU and 2GB memory were 117, 20, and 256 minutes for NIPS, 20News, and EachMovie data sets, respectively.

5. CONCLUSIONS

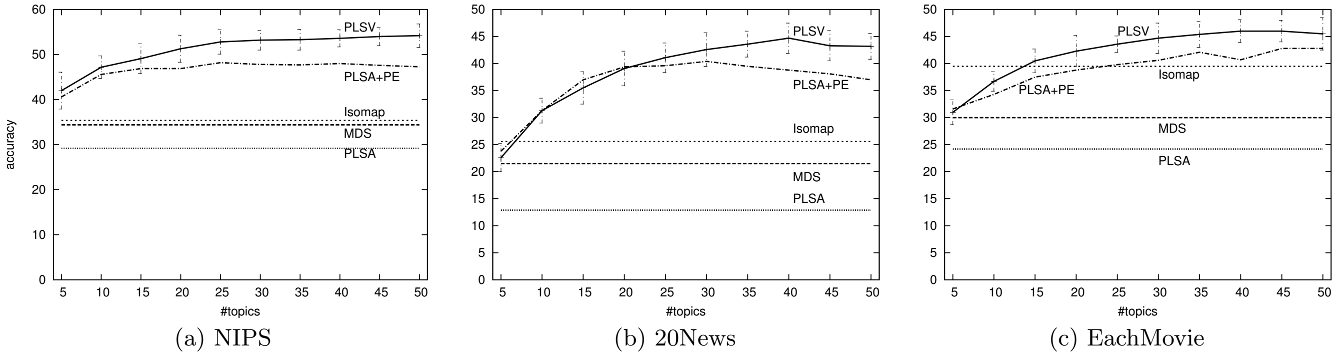
In this paper, we proposed a visualization method based on a topic model, Probabilistic Latent Semantic Visualization (PLSV), for discrete data such as documents. We have confirmed experimentally that PLSV can visualize documents with the latent topic structure. The results encourage us to believe that our data visualization approach based on PLSV is promising and will become a useful tool for visualizing documents.

Since PLSV is a probabilistic topic model, we can extend PLSV easily based on other research on topic models. Topic models have been proposed that model not only documents but also other information, for example images [2], time [21], authors [16] and citations [6]. We can also visualize this information with documents using the framework proposed



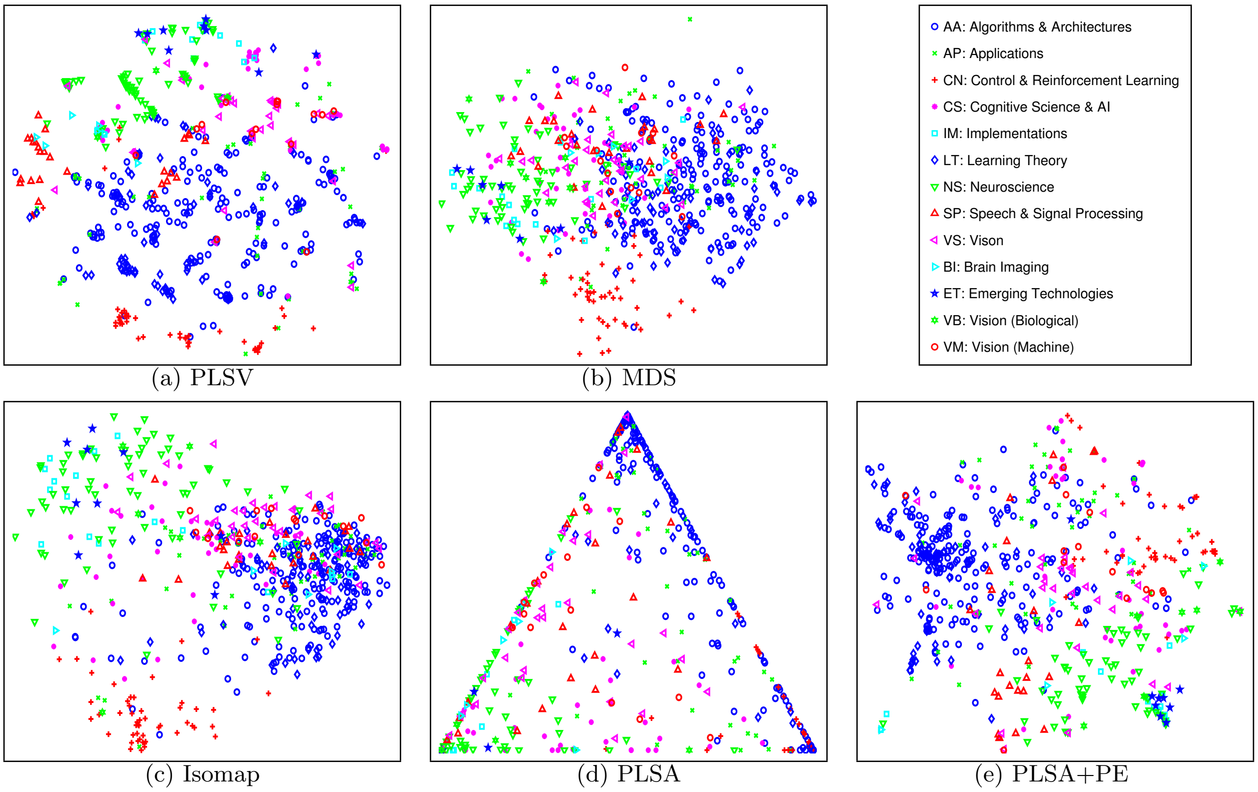
(a) NIPS (b) 20News (c) EachMovie

Figure 2: Accuracy with the k-nearest neighbor method in the visualization space with different numbers of neighbors k.



(a) NIPS (b) 20News (c) EachMovie

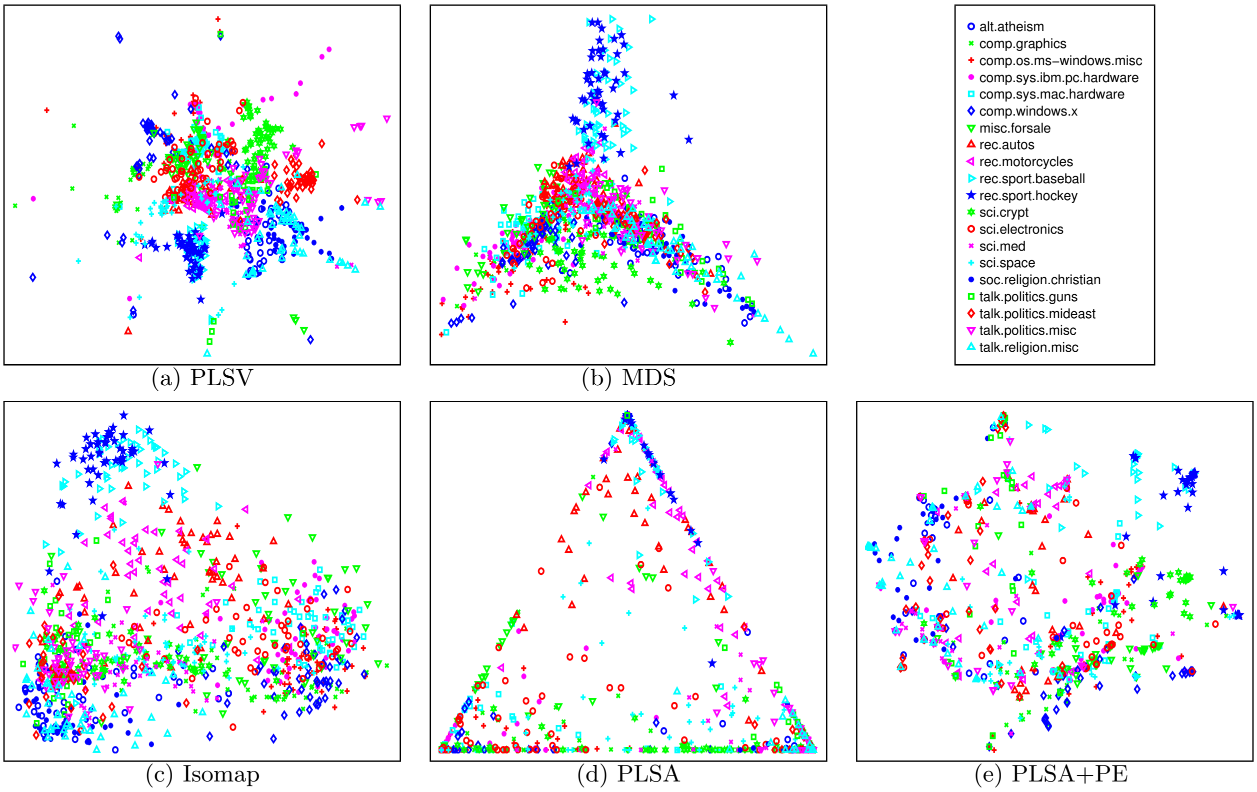
Figure 3: Accuracy with the one-nearest neighbor method in the visualization space with different numbers of topics Z for PLSV and PLSA+PE.



(a) PLSV (b) MDS

(c) Isomap (d) PLSA (e) PLSA+PE

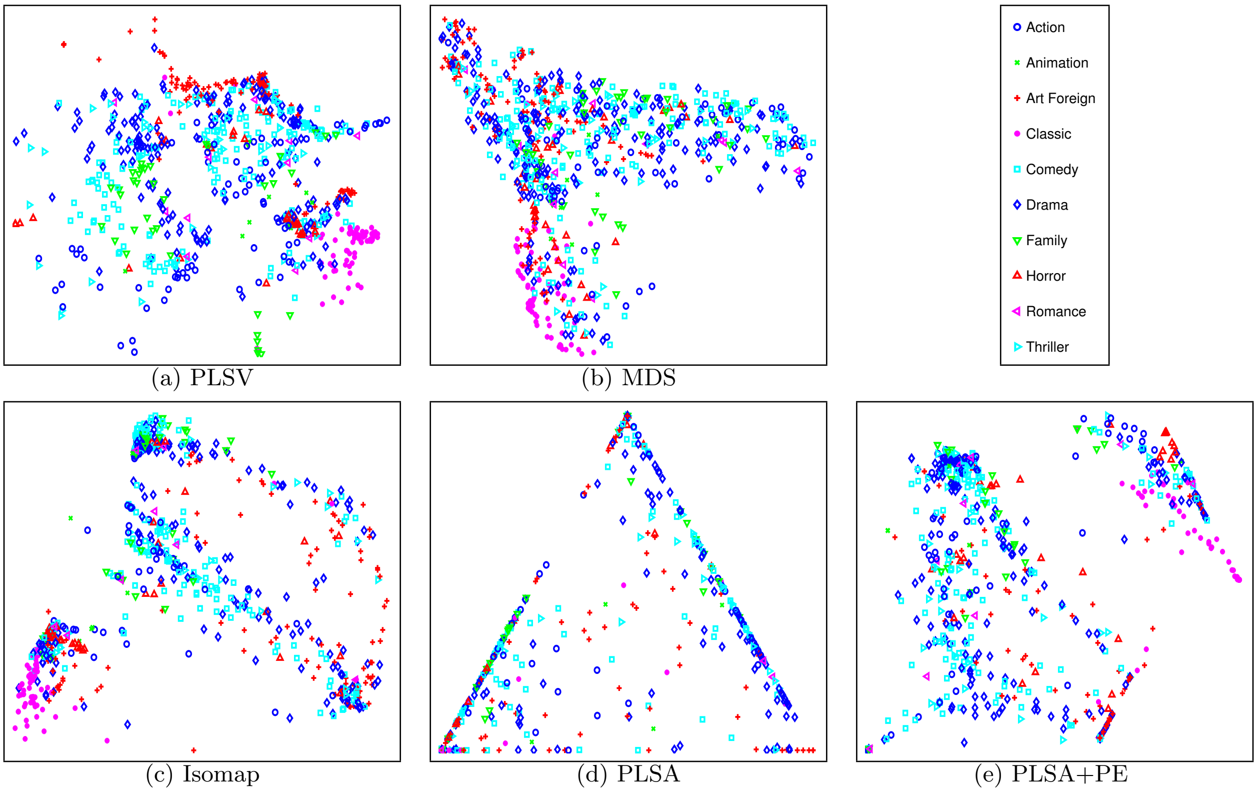
Figure 4: Visualization of documents in NIPS.



(a) PLSV (b) MDS

(c) Isomap (d) PLSA (e) PLSA+PE

Figure 5: Visualization of documents in 20News.



(a) PLSV (b) MDS

(c) Isomap (d) PLSA (e) PLSA+PE

Figure 6: Visualization of movies in EachMovie.

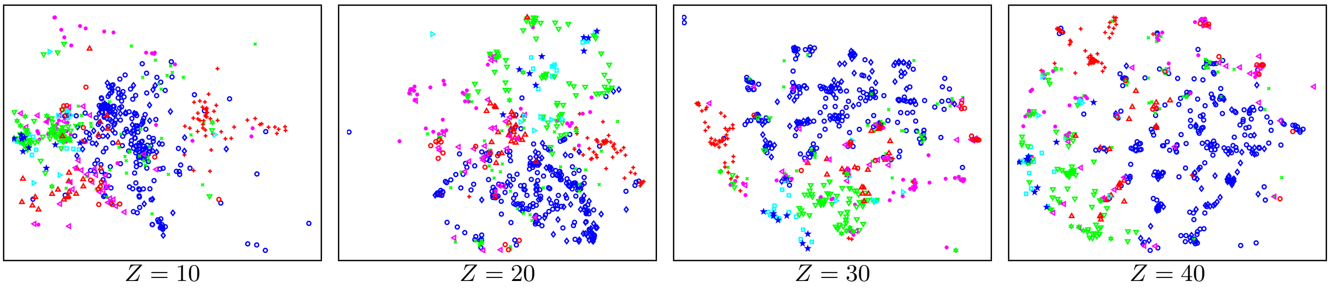
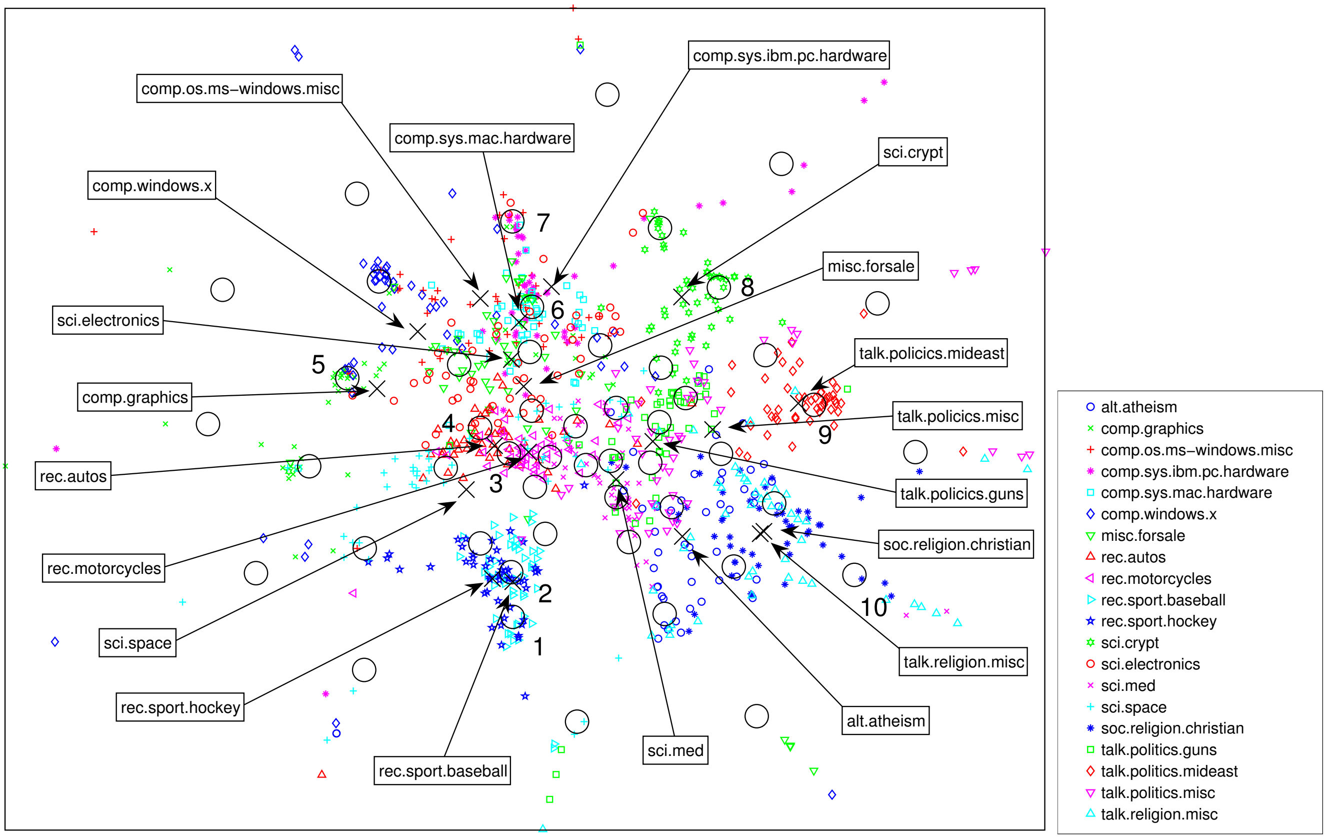


Figure 7: Visualization by PLSV with different numbers of topics Z on NIPS.



1	2	3	4	5	6	7	8	9	10
team	game	car	high	graphics	windows	scsi	clipper	israel	god
players	good	bike	engine	points	pc	printer	encryption	israeli	jesus
season	games	big	battery	lines	keyboard	bus	public	militia	ra
league	play	cars	car	point	drive	windows	system	arab	bible
nhl	baseball	water	stuff	line	mouse	drivers	government	jewish	christ
teams	guys	drive	low	reference	mac	hp	keys	jews	heaven
average	win	buy	bought	image	disk	speed	chip	turkish	people
hall	fans	thing	dealer	access	card	local	security	armenian	john
make	division	front	kind	program	memory	network	escrow	armenians	scripture
dave	guy	miles	speed	comp	dos	fonts	nsa	palestine	spirit

Figure 8: Visualization of documents in 20News by PLSV with $Z = 50$. Each point represents a document, and the shape and color represent the discussion group. Each black circle indicates a topic coordinate ϕ_z, and each black × indicates a label mean $\boldsymbol{\mu}_y$. The table at the bottom shows the ten most probable words for ten topics estimated with the visualization, i.e. the words are ordered according to $P(w|z, \boldsymbol{\Theta})$. The topic index corresponds to the number near the circle in the visualization.

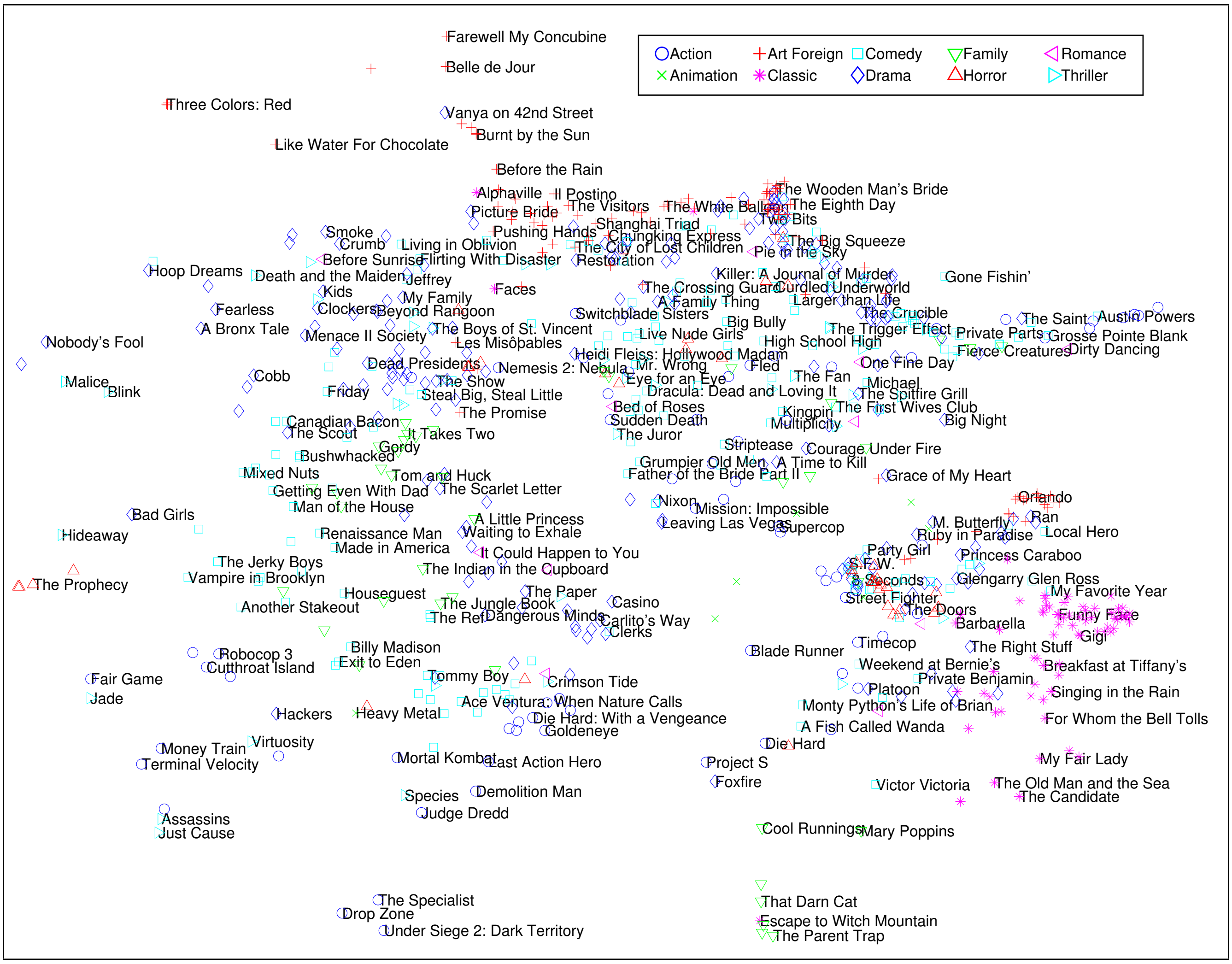


Figure 9: Visualization of movies in EachMovie by PLSV with $Z = 50$. Each point represents a movie, and the shape and color represent the genre. Some examples of movie titles are also shown.

in this paper. We assumed that the number of topics was known. We can automatically infer the number of topics by extending PLSV to a nonparametric Bayesian model such as the Dirichlet process mixture model [18]. In addition, we can achieve more robust estimation using the Bayesian approach, instead of MAP estimation, as in LDA [9].

6. REFERENCES

[1] C. M. Bishop, M. Svensen, and C. K. I. Williams. GTM: The generative topographic mapping. *Neural Computation*, 10(1):215–234, 1998.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th Annual International ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, 2003.

[3] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.

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

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

[6] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *ICML '07: Proceedings of the 24th International Conference on Machine Learning*, pages 233–240, 2007.

[7] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7:179–188, 1936.

[8] B. Fortuna, M. Grobelnik, and D. Mladenic. Visualization of text document corpus. *Informatica*, 29(4):497–502, 2005.

[9] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101 Suppl 1:5228–5235, 2004.

[10] T. Hofmann. Probabilistic latent semantic analysis. In *UAI '99: Proceedings of 15th Conference on Uncertainty in Artificial Intelligence*, pages 289–296, 1999.

[11] T. Iwata, K. Saito, N. Ueda, S. Stromsten, T. L. Griffiths, and J. B. Tenenbaum. Parametric embedding for class visualization. *Neural Computation*, 19(9):2536–2556, 2007.

[12] A. Kabán. Predictive modelling of heterogeneous sequence collections by topographic ordering of histories. *Machine Learning*, 68(1):63–95, 2007.

[13] A. Kabán, J. Sun, S. Raychaudhury, and L. Nolan. On class visualisation for high dimensional data: Exploring scientific data sets. In *DS '06: Proceedings of the 9th International Conference on Discovery Science*, 2006.

[14] K. Lang. NewsWeeder: learning to filter netnews. In *ICML '95: Proceedings of the 12th International Conference on Machine Learning*, pages 331–339, 1995.

[15] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Math. Programming*, 45(3):503–528, 1989.

[16] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *UAI '04: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494, 2004.

[17] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[18] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[19] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[20] W. Torgerson. *Theory and methods of scaling.* Wiley, New York, 1958.

[21] X. Wang and A. McCallum. Topics over time: a non-Markov continuous-time model of topical trends. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 424–433, 2006.

[22] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *INFOVIS '95: Proceedings of the 1995 IEEE Symposium on Information Visualization*, pages 51–58, 1995.