# Heterogeneous Topic Model by Non-negative Matrix Factorization


## Abstract

Topic models have been widely and successfully exploited for clustering. Nevertheless, it is necessary to know the form of probability distribution for most existent topic models, such as latent Dirichlet allocation (LDA). Also, those methods mainly operate on objects of a single type. In real applications, many tasks contain heterogeneous types of features (multiple views), e.g., stock mining with textual financial news and time-series price data. To attack these problems, a heterogeneous topic model (HTM) is proposed in terms of a graphical model representation. By avoiding knowing the form of probability distribution, the HTM is reduced to solving a non-negative matrix tri-factorization problem with certain constraints such that the proposed approach can be used for an arbitrary model. Here, two matrices represent topic mixtures in individual views and the other matrix measures the degree of the correlation of the topics from different sources. We have conducted experiments on one toy data and two real heterogeneous data sets (stock and text clustering). Experimental results validate the effectiveness of the proposed heterogeneous topic model.


## 1 Introduction

Topic models have been widely and successfully exploited for clustering problems by modeling a document mixture of topics. Nevertheless, most existent topic models are based on probabilistic latent semantic analysis (pLSA) [Hofmann, 1999b], latent Dirichlet allocation (LDA) [Blei *et al.*, 2002], and pachinko allocation model (PAM). Also, those methods operate on objects of a single type endowed with a measure of similarity or dissimilarity.

In recent years, literature also considers co-occurrence for heterogeneous objects with different feature types, such as [Blei and Jordan, 2003]. However, a Bayesian formulation of topic model requires knowing the specific probability distribution and the prior over topic mixtures, such as Dirichlet. In addition, the Bayesian integral over the hidden variables (usually in high dimensional space) makes it computationally infeasible. Although sampling methods, such as Gibbs sampling [Griffiths, 2002], can be adopted for handling the problem, the approximate Bayesian inference cannot be consistent in that it cannot recover the correct distribution.

To answer the problems, a heterogenous topic model is proposed in the paper to cluster data with different types of features, where one type of features is called as a view. Per view, "topics" can be generated by a mixture of "words" with a proportion. By expressing those in a graphical model, the correlation between two sets of topics from different sources together with the "word-topic" relations in each view are analogous to factors generated by non-negative matrix tri-factorization (NMTF) [Ding *et al.*, 2006].

In other words, the proposed HTM model can be solved via the technique used in non-negative matrix tri-factorization. Therefore, the correlation between data matrices (from different views) is factorized to three matrices, i.e., $\mathbf{P}$, $\mathbf{A}$, $\mathbf{Q}$, where $\mathbf{P}$ denotes the feature-topic proportion in one view, and $\mathbf{Q}$ is the feature-topic proportion in the other view. Accordingly, the elements in the two matrices satisfy the non-negative property. These two feature-topic matrices are analogous to the factors generated by the NMTF method. Different from the traditional NMTF method [Ding *et al.*, 2006], the topics in individual views (i.e., the columns of $\mathbf{P}$ and $\mathbf{Q}$) need not be orthogonal such that the correlation between topics in the same view can also be kept in the proposed method. Moreover, $\mathbf{A}$ captures the joint probability among the topics generated from different views on condition that the sum of all the elements in $\mathbf{A}$ is equal to one. In this regard, finer-grained clusters can be obtained using the correlation of two sets of topics generated from heterogeneous sources.

The novelties of the proposed approach are summarized as follows:(1) it is the first time to propose a heterogeneous topic model via non-negative matrix tri-factorization to handle the data from different sources; (2) it is not necessary to know the form of probability models of topics; (3) it can capture the correlation between the topics from different views. The proposed heterogeneous topic model is validated by NIPS papers from 1988 to 1999 with documents and abstracts and the Chinese A-shares with business scope descriptions and time-series price data. Experimental results show that better topic clusters can be captured by heterogeneous information through the correlation between two sets of topics from individual views.

The paper is organized as follows. Related work is given in Section 2. Section 3 describes the proposed heterogeneous topic model by non-negative matrix tri-factorization. Section 4 describes the data sets used and collected and reports experimental results. Finally, conclusion and extension are given in Section 5.

## 2 Related Work

For a single view task, the connection between topic models and non-negative matrix factorization has been widely established. In [Hofmann, 1999a], the topic model was firstly interpreted as matrix factorization by connecting the relation between latent semantic analysis (LSA, performing an SVD) [Furnas *et al.*, 1988] and probabilistic latent semantic analysis (PLSA) [Hofmann, 1999a] which has been successfully applied to different applications, such as document clustering. To handle discrete or positive only data, [B., 2002] argued that the topic models including probabilistic latent semantic analysis (pLSA) and latent Dirichlet allocation (LDA) and non-negative matrix factorization (NMF) [Lee *et al.*, 1999] are the instances of multinomial principal component analysis (PCA). Furthermore, [Gaussier and Goutte, 2005] showed that maximum likelihood solution of PLSA is equivalent to solving NMF with Kullback-Leibler (KL) divergence.

Usually in the text analysis application, the documents contain additional information such as author, title, venue, links, and others. There are few literature to incorporate those data to topic models. The author-topic model [Rosen-Zvi *et al.*, 2004] learns topics from both documents and authors by attaching authors to the topic proportions. The relational topic model uses additional linking information, such as documents linked by citation, webpage by hyperlinks [Chang and Blei, 2010], in which document is the same as the traditional one but the links between documents are assigned by the distance between their topic proportions. In [Wang *et al.*, 2007], a generalized component analysis (GCA) based topic model was proposed in terms of a two-layer factorization structure to capture words, authors and timestamps, where words are encoded as individual observations instead of word counts. Topic model is not only restricted for document analysis but also applied to computer vision by connecting the relations between images and taggings [Blei and Jordan, 2003].

Originally, the NMF is decomposed to two matrices with an orthogonal constraint for both matrices by enforcing a non-negative constraint for all the elements of two factor matrices [Lee *et al.*, 1999]. This is usually held for real applications, such as document clustering. Here, a document-term matrix is constructed with bag-of-word features. Then, the matrix is factored into a term-topic and a topic-document matrix, respectively. The topic-document matrix describes data clusters of related documents. However, the NMF is enforced by double orthogonal constraints for two factor matrices to obtain a unique solution, and also the low-rank approximation is rather poor. A 3-factor nonnegative matrix tri-factorization (NMTF) was proposed by [Ding *et al.*, 2006] to attack the problems. Initially, NMF and NMTF are exploited only for a single view data.

To handling heterogeneous data, the NMTF is further developed for two-view data analysis, in particular the co-clustering problem. [Wang *et al.*, 2011a] proposed a symmetric nonnegative matrix tri-factorization to co-clustering multi-type relational data considering both inter-type and intra-type relationships. [Zhang and Yeung, 2012] applied the NMTF to overlapping community detection, in which the two factors represent the membership of each node in each community and the remaining factor represents the interaction between the communities. Also, a fast version NMTF was proposed by constraining cluster indicator matrices instead of non-negative factor matrices [Wang *et al.*, 2011b]. Besides, the NMTF is also applied for classification on two-view data. [Li *et al.*, 2009] incorporated sentiment lexicon generated by experts to non-negative matrix tri-factorization for sentiment classification. [Shen and Li, 2011] selected both the "informative" examples (e.g., documents) and features (e.g., words) for active dual supervision.

## 3 Heterogeneous Topic Model via Non-Negative Matrix Factorization

Let the given data set make up of $n$ objects in one view $\mathbf{X} = (\mathbf{x}_i)_{i=1}^n, \mathbf{x} \in \mathcal{R}^p$ and in the other correlated view $\mathbf{Y} = (\mathbf{y}_j)_{j=1}^n, \mathbf{y} \in \mathcal{R}^q$. The aim of the paper is to find topic clusters provided by the information simultaneously from two views.

In the paper, an object is firstly deduced based on traditional topic models but in terms of heterogenous sources. However, it is necessary to define a prior distribution for such a two-view topic model, which prevents it from widely applying to a much complex problem. Also, the solution is calculated by the integral over the whole space. This makes it impossible in a high dimensional data without exploiting an approximation optimization technique. By introducing a matrix notation, the two-view topic model is reduced to a non-negative matrix tri-factorization problem by satisfying certain constraints.

### 3.1 Two-view Topic Models

In traditional topic models, two additional hidden random variables ($\mathbf{w}$, $\mathbf{z}$) that are unobserved are introduced to describe hidden topics, where $\mathbf{w} \in \mathcal{R}^s$ is a probability vector and a topic mixture for one view and $\mathbf{z} \in \mathcal{R}^t$ topic mixture for the other view.

Accordingly, the input vector $\mathbf{x}$ is generated from a mixture $\mathbf{w}$ over $s$ topics. Let $p_i(\mathbf{x})$ be the probability of topic $i$, then

$$p(\mathbf{x}|\mathbf{w}) = \sum_{i=1}^{s} \mathbf{w}_i p_i(\mathbf{x}). \tag{1}$$

Similarly, $\mathbf{y}$ is a mixture of $\mathbf{z}$ over $t$ topics. Let $p_j(\mathbf{y})$ be the probability of topic $j$, then

$$p(\mathbf{y}|\mathbf{z}) = \sum_{j=1}^{t} \mathbf{z}_j p_j(\mathbf{y}). \tag{2}$$

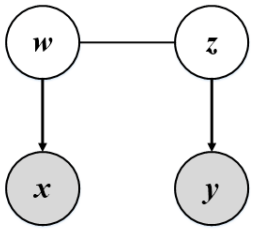



Figure 1: Two view topic model. Shaded random variables are observed from different sources.

| | |
|---|---|
| $\mathbf{x}, \mathbf{X}$ | object and object set in view one |
| $\mathbf{w}$ | hidden variable (topic) in view one |
| $\mathbf{P}$ | topic proportions in view one |
| $\mathbf{P}_i$ | the $i$-th row of $\mathbf{P}$ |
| $\mathbf{P}[i,:]$ | the $i$-th column of $\mathbf{P}$ |
| $\mathbf{y}, \mathbf{Y}$ | object and object set in view two |
| $\mathbf{z}$ | hidden variable (topic) in view two |
| $\mathbf{Q}$ | topic proportions in view one |
| $\mathbf{Q}[i,:]$ | the $i$-th column of $\mathbf{Q}$ |
| $\mathbf{Q}_i$ | the $i$-th row of $\mathbf{Q}$ |
| $\mathbf{A}$ | topic-topic correlation between two sets of objects from individual views |

Table 1: Notation used in the paper.

From (1), one can see that conditioned on the hidden variable $\mathbf{w}$, the input object $\mathbf{x}$ from one view is independent of the input vector $\mathbf{y}$ and its latent variable $\mathbf{z}$ from the other view. Likewise, $\mathbf{y}$ is also independent of $\mathbf{x}$ and $\mathbf{w}$ conditioned on $\mathbf{z}$. However, two mixtures from different sources should be correlated. Based on this rationale, a graphical model representation is given in Figure 1. Then, we have

$$\begin{aligned} E_{\mathbf{x},\mathbf{y}}\mathbf{x}\mathbf{y}^\top &= \iint \mathbf{x}\mathbf{y}^\top p(\mathbf{x},\mathbf{y})d\mathbf{x}d\mathbf{y} \\ &= \iint \mathbf{x}\mathbf{y}^\top \sum_{\mathbf{w},\mathbf{z}} p(\mathbf{x}|\mathbf{w})p(\mathbf{y}|\mathbf{z})p(\mathbf{w},\mathbf{z})d\mathbf{x}d\mathbf{y} \\ &= \sum_{i,j} E_{\mathbf{w},\mathbf{z}}\mathbf{w}_i\mathbf{z}_j \int \mathbf{x}p_i(\mathbf{x})d\mathbf{x} \int \mathbf{y}^\top p_j(\mathbf{y})d\mathbf{y}. \end{aligned} \tag{3}$$

Let $\mathbf{P} \in \mathcal{R}^{p\times s}$ be a matrix, where the $i$-th column $\mathbf{P}[i,:] = \int \mathbf{x}p_i(\mathbf{x})d\mathbf{x}$. Let $\mathbf{Q} \in \mathcal{R}^{q\times t}$ be a matrix, where the $j$-th column $\mathbf{Q}[j,:] = \int \mathbf{y}p_j(\mathbf{y})d\mathbf{y}$. Let $\mathbf{A} \in \mathcal{R}^{s\times t}$ be the topic correlation matrix, where $\mathbf{A}_{i,j} = E_{\mathbf{w},\mathbf{z}}\mathbf{w}_i\mathbf{z}_j$. Then, (3) can be rewritten in a matrix notation as

$$E_{\mathbf{x},\mathbf{y}}\mathbf{x}\mathbf{y}^\top = \mathbf{P}\mathbf{A}\mathbf{Q}^\top. \tag{4}$$

Notation used in the paper is shown in Table 1.

### 3.2 The Proposed Heterogeneous Topic Model

To obtain topic clusters simultaneously using heterogeneous object features, the aim of our work is reduced to finding three matrices, by expressing topic mixtures from individual views and topic-topic correlations between two views based on (4).

In this case, the matrix $\mathbf{P} \in \mathcal{R}^{p\times s}$ denotes the word-topic proportions with $s$ topics generated from one view, e.g., a document being a mixture of topics. Then, the elements in $\mathbf{P}$ should be non-negative, i.e., $\mathbf{P}[i,j] \geq 0$ and also, the matrix should satisfy $\sum \mathbf{P}[i,:] = 1$, where $\mathbf{P}[i,:]$ denotes the $i$-th column of $\mathbf{P}$. Similarly, the same conditions should be held for the matrix $\mathbf{Q} \in \mathcal{R}^{q\times t}$ denoting the word-topic proportions with $t$ topics generated from the other correlated view. The correlation matrix $\mathbf{A} \in \mathcal{R}^{s\times t}$ between two sets of topics generated from two views should satisfy the conditions $\mathbf{A}[i,j] \geq 0$ and $\sum_{i,j} \mathbf{A}[i,j] = 1$.

To fulfil the target with the constraints aforementioned, we can find the matrices by

$$E_{\mathbf{x}\mathbf{y}}\mathbf{x}\mathbf{y}^\top = \mathbf{P}\mathbf{A}\mathbf{Q}^\top \tag{5}$$

$$s.t.\ \ \mathbf{P},\mathbf{A},\mathbf{Q} \geq 0 \tag{6}$$

$$||\mathbf{P}[i,:]||_1 = 1, 1 \leq i \leq s \tag{7}$$

$$||\mathbf{Q}[j,:]||_1 = 1, 1 \leq j \leq t \tag{8}$$

$$||\mathbf{A}||_1 = 1. \tag{9}$$

### 3.3 Numerical solution

From (5), we can obtain for each $1 \leq i \leq p$ and $1 \leq j \leq q$

$$E_{\mathbf{x}^i\mathbf{y}^j}\mathbf{x}^i\mathbf{y}^j = \mathbf{P}_i^\top \mathbf{A}\mathbf{Q}_j \tag{10}$$

where $\mathbf{P}_i^\top$ is the $i$-th row of $\mathbf{P}$, and $\mathbf{Q}_j^\top$ is the $j$-th row of $\mathbf{Q}$ with the constraints from (6) to (9). To solve the optimization in (10), we can find the three matrices by the following regression problem

$$[\mathbf{P},\mathbf{A},\mathbf{Q}] = \arg\min_{\mathbf{P},\mathbf{A},\mathbf{Q}} \sum_{i,j} \left(\mathbf{P}_i^\top \mathbf{A}\mathbf{Q}_j - E_{\mathbf{x}^i\mathbf{y}^j}\mathbf{x}^i\mathbf{y}^j\right)^2 \tag{11}$$

with the constraints from (6) to (9). Then, an alternative least squares can be applied to find $\mathbf{P}, \mathbf{A}, \mathbf{Q}$ by fixing two matrices to solve the remaining matrix in an iterative way, namely,

- Fixing $\mathbf{P}$ and $\mathbf{Q}$ to solve $\mathbf{A}$;
- Fixing $\mathbf{P}$ and $\mathbf{A}$ to solve $\mathbf{Q}$;
- Fixing $\mathbf{A}$ and $\mathbf{Q}$ to solve $\mathbf{P}$.

For the whole data set $\mathbf{X} = (\mathbf{x}_i)_{i=1}^n, \mathbf{x} \in \mathcal{R}^p$ in the one view and $\mathbf{Y} = (\mathbf{y}_j)_{j=1}^n, \mathbf{y} \in \mathcal{R}^q$ in the other correlated view, with the matrix notation we have

$$\begin{aligned} E_{\mathbf{X}\mathbf{Y}}\mathbf{X}\mathbf{Y}^\top &= E\left(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_n\right)\left(\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_n\right)^\top \\ &= \sum_i E_{\mathbf{x}_i\mathbf{y}_i}\mathbf{x}_i\mathbf{y}_i^\top. \end{aligned}$$

Therefore, we have

$$E_{\mathbf{x}\mathbf{y}}\mathbf{x}\mathbf{y}^\top = \frac{1}{n}E_{\mathbf{X}\mathbf{Y}}\mathbf{X}\mathbf{Y}^\top.$$

## 4 Experiments

In the paper, a heterogeneous topic model by non-negative matrix tri-factorization is proposed to cluster data objects from different sources. The proposed approach is validated by two real data sets, namely, NIPS papers and the Chinese A-shares corpus. In order to better analyze the results, a subset of the stock data was generated as a toy data for visualization.


Figure 2: Daily candlesticks chart for Shanghai Composite Index from May 4 to Dec. 31, 2012.

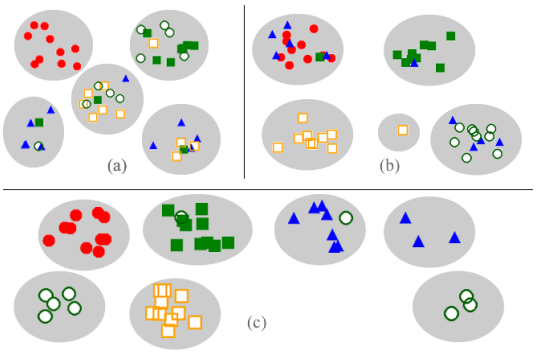



Figure 3: Illustration of topic clusters learnt by (a) the one-view BSD corpus by the LDA, (b) the one-view time-series price data by the Kmeans and (c) heterogeneous data with both the BSD corpus and the time-series price data using the proposed HTM. Shaded Circle: Banking; Shaded Square: Iron and Steel; Shaded Triangle: Pharmaceuticals; Hollow Square: Alcohol; Hollow Circle: Information Technology. .

### 4.1 Data Sets

**NIPS13**: The data set contains papers from the NIPS conferences between 1987 and 1999. The conference is characterized by contributions from a number of different research communities in the general area of learning algorithms. The original collection of NIPS papers [1] contains 1,740 papers with a total of 2,301,375 word tokens and a vocabulary size of 13,649 unique words. All documents labeled as 'unrecognized' were removed for better evaluation. Accordingly, there are 1641 documents with the vocabulary size of 6417 words. In the following experiments, the one view contains the whole papers and the other view consists of the corresponding abstracts with 48 topics.

**Stock2209**: The data were collected from the Chinese A-shares market. The one view is representative and clean data, i.e., the lines of business (or business scope description, BSD). Here, the 2,209 BSD documents are exploited which contain textual scope descriptions and lists of external parties (customer, supplier, partners) and their roles (external actors). The features are extracted by sending the BSD corpus to the language processing pipeline, including tokenization, Part-of-Speech (PoS) tagging. Then, the words in the stop list and the words which occur less than and equal to two times are dropped. Then, the feature representation is obtained based on term frequency (TF) by the *bag-of-words* assumption.

The other view is time-series price data. To check the influence of a general trend, the price sets made up of severe concussion were selected according to the benchmark Shanghai Composite Index (SCI) as shown in Figure 2. In the first period $T_1$ from May 4 to Sept. 26, 2012, the SCI continues to drop as a downward trend for more than 2 months (the period shown as "1" in the figure). Then, there is an upturn for two weeks (the period shown as "2" in the figure), and again drops for more than one month (the period shown as "3" in the figure). Finally, a sharp rise comes after the second fall (the period shown as "4" in the figure). The periods "2", "3" and "4" consist of the second period ($T_2$) from Sept. 27 to Dec. 31, 2012.

Although an arbitrary model can be used for the proposed algorithm, the price data are processed to have a TF meaning in order to use the LDA model as an initialization. Namely, a 2T-dimensional vector is constructed with the first T elements denoting the increasing rate and the remaining T ones describing a increasing/decreasing rate in two successive days. If the rate of change is not an integer, it is rounded off. The maximum value is ten as the Chinese stock market regulators impose the value ten as daily price limits on individual stock price movements.

**Stock50**: In order to better analyze the results provided by the proposed approach, 50 stocks containing 5 industries from the **Stock2209** were selected as a toy data. The 5 industries are Banking, Pharmaceuticals, Information Technology, Iron and Steel, and Alcohol, respectively and ten stocks from each industry were randomly selected conditioned on that those stocks got listed on the Chinese A-shares market before May 4, 2012. The data descriptions from two views are the same as the **Stock2209**.

### 4.2 Experimental results

According to semi-strong-form efficient market hypothesis (EMH) [Fama, 1970], stock prices reflect all public information. Accordingly, a Pearson correlation coefficient (PCC) [Stigler, 1989] is exploited to reflect the correlation between two stocks in time-series price data in a fixed time period. Furthermore, the p-value is adopted to show the relationship between two variables. To evaluate the correlation among stocks in a stock cluster, an average PCC (aPCC) and an average p-value over all the stocks belonging to a same cluster are reported in the end.

NIPS13 is a document clustering problem and the groundtruth clusters are highly overlapped, such as "Speech, Handwriting and Signal Processing ". Therefore, the topic-topic correlation matrix is printed and visualized in the experiments.

As the topic proportion matrices $\mathbf{P}$ and $\mathbf{Q}$ should be firstly initialized by a clustering algorithm. In the paper, the traditional topic model LDA, Kmeans clustering, and the tra-

[1] Available at http://www.cs.toronto.edu/ roweis/data.html.

| | $T_1$ | | $T_2$ | | $T_1 + T_2$ | |
|---|---|---|---|---|---|---|
| | aPCC | p-value | aPCC | p-value | aPCC | p-value |
| Expert | 0.3481 | 0.0521 | 0.4892 | 0.0548 | 0.3934 | 0.0428 |
| LDA | 0.3288 | 0.0533 | 0.4658 | 0.0599 | 0.3723 | 0.0437 |
| Kmeans | 0.3410 | 0.0452 | 0.5360 | 0.0473 | 0.4008 | 0.0371 |
| NMF | 0.3702 | 0.0458 | 0.5298 | 0.0478 | 0.4218 | 0.0378 |

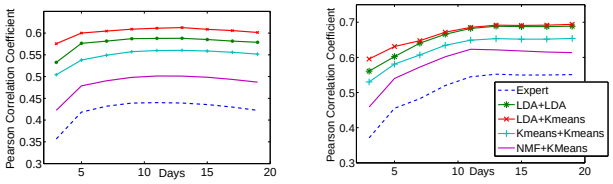
Table 2: Average Pearson correlation coefficients (aPCC) and average p-value (p-value) over the time period $T_1$, the time period $T_2$ , and two time periods $T_1 + T_2$ together learnt by different approaches including LDA, Kmeans, and NMF on the one-view stock BSD corpus for Stock2209.

ditional non-negative matrix factorization are adopted in the following.

**Stock50: a Toy data**

A visualization result on the 50 stocks in 5 sectors (topic clusters) assigned by domain experts is shown in Figure 3. Figure 3(a) shows 5 topic clusters learnt by the LDA on the single view BSD corpus, Figure 3(b) is provided by the Kmeans on the other view price data, and Figure 3(c) is the topic clusters by the proposed HTM on the two types of sources, i.e., the BSD and Price data. It is easy to see that the proposed approach obtains finer clustering information by considering not only business scopes but also the similar trends of stock prices. If only a single view, i.e., the BSD data, is exploited to generate sectors, the groups cannot reflect the stock ups and downs behaviors in the same sector as the BSD data is kept fixed usually after the companies were initial public offering. Nevertheless, the stock groups can be also defined by both one view, the BSD and the other view such as price data. Then, a sector generated contains the shares not only from the same industry but also with a similarly up or down trend. This is particularly useful to tracking "sector rotation" where if a bellwether leading the other stocks in the same sector sharply rises, it is of high probability for the other stocks in the same sector to increase afterward.

**Stock2209**




(a) (b)

Figure 4: The average Pearson correlation coefficients over different lengths of days assigned by human experts and learnt by the proposed HTM model with different initial approaches on $\mathbf{P}$ and $\mathbf{Q}$, i.e., LDA+LDA, LDA+Kmeans, Kmeans+Kmeans, and NMF+Kmeans, respectively on two time periods, i.e. (a) $T_1$ and (b) $T_2$.

The average Pearson correlation coefficients and average p-values are shown in Table 2 on a single-view data and in Table 3 on two-view data. The results in Table 2 are grouped by human expert and by different clustering approaches on a single view, i.e., the BSD corpus. Except for the results by the LDA, the linear correlations among a same cluster over the first time period $T_1$ learnt by a single view BSD corpus with different methods are slightly better than those by the Experts. Similarly, the predicted results on the average PCCs and the average p-values over the second time period ($T_2$) and the whole time period ($T_1 + T_2$) are better than the Experts.

| Initial Methods | | $T_1$ | | $T_2$ | | $T_1 + T_2$ | |
|---|---|---|---|---|---|---|---|
| BSD($\mathbf{P}$) | Price($\mathbf{Q}$) | aPCC | p-value | aPCC | p-value | aPCC | p-value |
| LDA | LDA | 0.5418 | 0.0349 | 0.6394 | 0.0404 | 0.5794 | 0.0285 |
| LDA | Kmeans | 0.5448 | 0.0340 | 0.6523 | 0.0442 | 0.5876 | 0.0249 |
| LDA | NMF | 0.5555 | 0.0288 | 0.6072 | 0.0509 | 0.5626 | 0.0334 |
| Kmeans | LDA | 0.2889 | 0.0526 | 0.4358 | 0.0637 | 0.3407 | 0.0455 |
| Kmeans | Kmeans | **0.5788** | **0.0320** | 0.5836 | 0.0602 | 0.5720 | 0.0242 |
| Kmeans | NMF | 0.4192 | 0.0386 | 0.4301 | 0.0664 | 0.4006 | 0.0404 |
| NMF | LDA | 0.4496 | 0.0652 | 0.5823 | 0.0458 | 0.4810 | 0.0796 |
| NMF | Kmeans | 0.4962 | 0.0273 | 0.5365 | 0.0508 | 0.5114 | 0.0303 |
| NMF | NMF | 0.5471 | 0.0320 | 0.6131 | 0.0416 | 0.5577 | 0.0258 |

Table 3: Average Pearson correlation coefficients (aPCC) and average p-value (p-value) over the time period $T_1$, the time period $T_2$, and two time periods $T_1 + T_2$ together learnt by the proposed HTM with different initial approaches including LDA, pLSA, Kmeans, and NMF on the BSD corpus and the time-series price data on $T_1$ for Stock2209.

The aPCCs and p-values by the proposed approach HTM on both the BSD corpus and the price data are shown in Table 3. The stock prices over the time period $T_1$ together with the BSD corpus were used to train the HTM model. By incorporating the price data to the model, the correlations among stocks in a same topic cluster are significantly better than those by the Experts and by a single view data according to the average Pearson correlation coefficients as well as p-values, e.g., aPCCs and p-values are 0.5788 and 0.0320, much better than 0.3481 and 0.0521 by the Experts, where a low p-value means a statistically significant relationship between two variables.

In order to check the efficiency of the topic clusters (or sectors) generated by the HTM model, the average Pearson correlation coefficients over different lengths of days are reported in Figure 4 by Experts and by the proposed models with different initial methods on two-view topic proportion matrices $\mathbf{P}$ and $\mathbf{Q}$, such as LDA+LDA, LDA+Kmeans, Kmeans+Kmeans, and NMF+Kmeans, respectively. Figure 4(a) shows the results on time period $T_1$ while Figure 4(b) are the predicted results on time period $T_2$. The correlations within different time periods by the proposed model are significantly and consistently increased than those by human experts. The similar results are obtained for the prediction results on time $T_2$.

**NIPS13**

According to the description of NIPS13 corpus, 48 topic clusters in each view are initialized at the beginning. After learnt

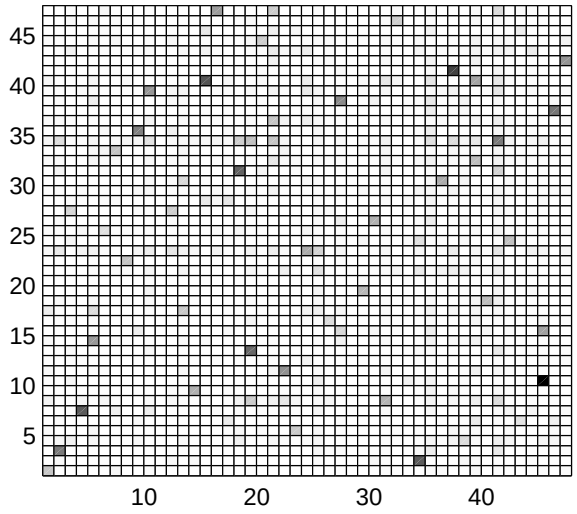



Figure 5: Topic correlation matrix $\mathbf{A}$ by the the proposed HTM model on the NIPS13 corpus with the two views, i.e., the full contents of papers and the abstracts.

by the HTM, two sets of topics in individual views are generated as well as the correlations of topics. Figure 5 show the visualization the topic-topic correlation probabilities, where each grid denotes a correlation probability of two topics from individual views. Here, colors in the grids represents the degrees of co-occurrence of two topics: the darker color is, the higher correlation probability is. Due to space limitation, only topics with the highest co-occurrence probability, i.e., the 10th topic in the full-text view and the 45th topic in the abstract view are shown in Table 4. From the descriptions, the semantic meaning of two topics can be inferred, such as "Neuroscience". The initial topic proportions of $\mathbf{P}$ and $\mathbf{Q}$ are initialized by the LDA for simplicity.

| Full-Text | | Abstracts | |
|---|---|---|---|
| Topic 10 | | Topic 45 | |
| Words | Prob. | Words | Prob. |
| input | 0.013 | patterns | 0.013 |
| neurons | 0.011 | cells | 0.013 |
| patterns | 0.011 | activity | 0.01 |
| synaptic | 0.011 | input | 0.009 |
| activity | 0.01 | model | 0.009 |
| pattern | 0.01 | synaptic | 0.009 |
| connections | 0.007 | pattern | 0.009 |
| figure | 0.006 | cortical | 0.007 |
| synapses | 0.006 | cell | 0.007 |
| neuron | 0.005 | cortex | 0.007 |

Table 4: Topics with the highest co-occurrence probability in $\mathbf{A}$, i.e., the 10th topic in the full-text view and the 45th topic in the abstract view.

## 5 Conclusion

In the paper, a heterogeneous topic model has been developed to find topic clusters by simultaneously utilizing data information from different sources. By modeling the correlation in a graphical model, the objective of two-view topic model is reduced to non-negative matrix tri-factorization with special constraints. Here, the topic proportions from each view is assumed as a factor matrix and the topic-topic correlations from heterogeneous data types is as another matrix. Namely, the correlation between data matrices (from different views) is factorized to three matrices, i.e., $\mathbf{P}$, $\mathbf{A}$, $\mathbf{Q}$. In the algorithm, $\mathbf{P}$ denotes the "word"-topic proportion in one view, and $\mathbf{Q}$ denotes the "word"-topic proportion in the other view. In addition, the elements in the two matrices satisfy the non-negative property. The topic-topic correlation matrix $\mathbf{A}$ denotes the joint probability among the topics generated from different views on condition that the sum of all the elements in $\mathbf{A}$ is equal to one.

Therefore, it is not necessary for the proposed method to know the concrete form of the probability model $\mathbf{P}_i(\mathbf{x})$ and $\mathbf{Q}_j(\mathbf{y})$. In other words, the method works for arbitrary models without the need to know the form of probability distribution. In comparison, a Bayesian formulation of topic models requires knowing the specific probability distribution such as Dirichlet.The other advantage of the proposed HTM model is that the solution of the proposed approach is consistent when the sample size approaches infinity such as in Big Data era while LDA-like models are inconsistent as an exact inference over the whole hidden variable spaces is infeasible. When documents have small number of words, the problem becomes much severe.

The proposed HTM model has been evaluated on NIPS13 corpus and the Chinese A-shares with the business scope description (in texts) and time series price data. Experimental results show that (1) co-occurrence topics in individual views can be identified by high probabilities from correlation probability matrix $\mathbf{A}$; (2) significantly higher correlation among stocks in a same sector (topic cluster) can be obtained compared to that assigned by human expert and provided by traditional clustering results on the information of a single view.

In the proposed HTM, the initialization of the proportional matrices has a significant influence on topic clustering results. This should be further investigated as a future development. In real applications, data usually contain more than 2 views, such as a paper with linking information, authors, venues, etc. The extension to more than two views will be further studied as another future development.

## References

[B., 2002] Wray L. B. Variational extensions to em and multinomial pca. In *Proceedings of the 13th European Conference on Machine Learning*, ECML '02, pages 23–34, 2002.

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

[Blei *et al.*, 2002] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 2002.

[Chang and Blei, 2010] J. Chang and D. Blei. Hierarchical relational models for document networks. *Ann. Appl. Stat.*, 4(1):124–150, 2010.

[Ding *et al.*, 2006] C. Ding, T. Li, W. Peng, and H. Park. Orthogonal nonnegative matrix t-factorizations for clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 126–135, 2006.

[Fama, 1970] E. F. Fama. Efficient capital markets: A review of theory and empirical work. *Journal of Finance*, 25(2):383C417, May 1970.

[Furnas *et al.*, 1988] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R. A. Harshman, L. A. Streeter, and K. E. Lochbaum. Information retrieval using a singular value decomposition model of latent semantic structure. In *Proceedings of the 11th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 465–480, 1988.

[Gaussier and Goutte, 2005] E. Gaussier and C. Goutte. Relation between plsa and nmf and implications. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '05, pages 601–602, 2005.

[Griffiths, 2002] T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. *Standford University*, 518(11):1–3, 2002.

[Hofmann, 1999a] T. Hofmann. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, UAI'99, pages 289–296, 1999.

[Hofmann, 1999b] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57, 1999.

[Lee *et al.*, 1999] D.D. Lee, H.S. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[Li *et al.*, 2009] T. Li, Y. Zhang, and V. Sindhwani. A nonnegative matrix tri-factorization approach to sentiment classification with lexical prior knowledge. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 244–252, 2009.

[Rosen-Zvi *et al.*, 2004] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, 2004.

[Shen and Li, 2011] C. Shen and T. Li. A non-negative matrix factorization based approach for active dual supervision from document and word labels. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 949–958, 2011.

[Stigler, 1989] S. M. Stigler. Francis galton's account of the invention of correlation. *Statistical Science*, 4(2):73C79, 1989.

[Wang *et al.*, 2007] X. Wang, C. Pal, and A. McCallum. Generalized component analysis for text with heterogeneous attributes. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 794–803, 2007.

[Wang *et al.*, 2011a] H. Wang, H. Huang, and C. Ding. Simultaneous clustering of multi-type relational data via symmetric nonnegative matrix tri-factorization. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 279–284, 2011.

[Wang *et al.*, 2011b] H. Wang, F. Nie, H. Huang, and F. Makedon. Fast nonnegative matrix tri-factorization for large-scale data co-clustering. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Two*, pages 1553–1558, 2011.

[Zhang and Yeung, 2012] Y. Zhang and D. Yeung. Overlapping community detection via bounded nonnegative matrix tri-factorization. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 606–614, 2012.