# Feature Sampling Based Unsupervised Semantic Clustering for Real Web Multi-View Content

**Xiaolong Gong, Linpeng Huang, Fuwei Wang**
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
{gxl121438, lphuang, wfwzy2012}@sjtu.edu.cn

**Abstract**

Real web datasets are often associated with multiple views such as long and short commentaries, users preference and so on. However, with the rapid growth of user generated texts, each view of the dataset has a large feature space and leads to the computational challenge during matrix decomposition process. In this paper, we propose a novel multi-view clustering algorithm based on the non-negative matrix factorization that attempts to use feature sampling strategy in order to reduce the complexity during the iteration process. In particular, our method exploits unsupervised semantic information in the learning process to capture the intrinsic similarity through a graph regularization. Moreover, we use *Hilbert Schmidt Independence Criterion* (HSIC) to explore the unsupervised semantic diversity information among multi-view contents of one web item. The overall objective is to minimize the loss function of multi-view non-negative matrix factorization that combines with an intra-semantic similarity graph regularizer and an inter-semantic diversity term. Compared with some state-of-the-art methods, we demonstrate the effectiveness of our proposed method on a large real-world dataset $Doucom$ and the other three smaller datasets.

## 1 Introduction

Real world application has to deal with data presenting various perspectives and is usually characterized by various heterogeneous sources of information. Web pages, multimedia documents, user profiles are examples of data that can be organized into multi-graphs. It is common that relevant multi-view data from different sources may have semantic correlations. For example, a movie item in IMDB often consists of many user reviews and one plot summary; An artist in Last.fm is always associated with its description and lots of short comments. As different semantic information may emphasize different aspects of the data, it gives rise to an emerging demand to explore the interactions between multi-view data for the application like multi-view clustering (Bickel and Scheffe 2004; Chaudhuri et al. 2009; Liu et al. 2013), co-regularized spectral clustering (Kumar, Rai, and Daumé 2011; Kumar and Daumé 2011) and cross-view retrieval (Zhai et al. 2013). Different from the traditional data with a single view, large-scale multi-view web contents commonly have the following properties: 1) Every single view of one web item[1] has its own feature sets. 2) Different views of one item share some consistency in semantic information. Just like a new story may be reported by different news resources, but the underlying content will not be changed. 3) The diversity information among the representations of one web item exists in multi-view data, and each representation corresponds to a single view content. Taking our $Doucom$ dataset as an example, a movie item in the web community consisting of multiple views including summary, short comment, long review, and user group. And each view of a movie item mainly contains specific information in the relevant content.

According to the above properties, multi-view clustering has attracted more and more attention because it can explore the intrinsic structure of the multi-view dataset and handle large numbers of unlabeled data. Generally, the main challenge lies in how to make use of the complementary characteristics embedded in the multiple sources of information and exploit the interactions and correlations between a various number of views accurately and automatically. Plenty of multi-view clustering algorithms have been developed to solve this problem. Some methods perform multi-view clustering through merging the clustering results from different individual views (Long, Yu, and Zhang 2008; Greene and Cunningham 2009). Some works (Cai, Nie, and Huang 2013; Cheng et al. 2013; J. Sun and Kranzler 2014) seek groupings that are consistent across different views. Recently, some researches (Liu et al. 2013; He et al. 2014) focus on nonnegative matrix factorization(NMF) framework, which attempts to interpret the distinction between different views via pairwise representation. All the above studies have shown effective in multi-view clustering, but they suffer from some inevitable problems: 1) Most of the studies ignore the curse of dimensionality in feature space with the growth of data size. This problem leads to a significant computational challenge when optimization algorithms load complete view matrix with large dimensions. Most recently proposed multi-view clustering methods (Hoyer 2004; Mohammadiha and Leijon 2009; Sun et al. 2015; Gong, Wang, and Huang 2017) utilized complete feature space to deal with the above problem,



[1] In this paper, we use "item" and "instance" interchangeably

whose optimization ignores the computational complexity. For instance, alternate minimization algorithm should be iterated hundreds of times for each original view matrix observation. 2) Some previous methods do not consider the unsupervised semantic similarity and diversity information among data instances in latent representation. They used supervised label information as similarity measurement to capture semantic information. Especially, the manual labels are insufficient in real large web data. The multiple clustering solutions can be accurately achieved if we explore the diversity information. For example, some movies directed by the same director could be grouped into one cluster according to the summary view and some movies that have different genres could possibly be grouped into one cluster due to the detailed description and profound implication from the view of the long review.

In this paper, we propose a novel multi-view unsupervised semantic clustering method, dubbed the name Feature Sampled Unsupervised Semantic Clustering ($FSUSC$) for real web multi-view content. We use graph regularization to capture the intrinsic intra-semantic similarity information. We also use the *Hilbert Schmidt Independence Criterion* (HSIC)(Arthur et al. 2005) as a co-regularizer term to enforce the inter-semantic diversity of the jointly learned representations. Although a few existing methods (Wang et al. 2015; Zhu et al. 2013; Zheng et al. 2011; Yang et al. 2014) apply graph regularization to various applications, they only observe the original data information, which motivates us to explore the semantic structure information and identify a common low-dimensional space of the data instance across multiple views in our work, thus reducing the memory consumption if data refers to a large view numbers. We formulate our multi-view clustering as a joint optimization problem that minimizes the reconstruction errors over the multiple views. The contributions of this paper can be summarized as the following:

- We construct a large-scale organized dataset from web community, namely **Doucom**, where feature space is much larger than other previous real-world datasets. The proposed method explicitly reduces the computational complexity by performing feature sampling at each iteration in four real web datasets.
- $FSUSC$ leverages unsupervised semantic information to refine the graph regularizer during the step of graph construction and uses HISC to measure diversity among data representations.
- To solve the objective function of $FSUSC$, we derive a new iterative updating optimization scheme and our proposed method can achieve state-of-the-art results in terms of accuracy and normalized mutual information.

## 2 Feature Sampled Unsupervised Semantic Clustering ($FSUSC$)

### Problem Statement

Before we describe the formulation of the problem, we summarize some notations used in this paper in Table 1. Let $l = 1, \cdots, n_v$, an arbitrary original data matrix in view $l$ is $X^{(l)} = [x_1^{(l)}, x_2^{(l)}, \cdots, x_m^{(l)}]^T \in \mathbb{R}_+^{m \times d^{(l)}}$, where each row ${x_i^{(l)}}^T (1 \leq i \leq m)$ represents a data instance and each column represents one feature. Non-negative matrix factorization (Lee and Seung 2001) aims to find two factors with non-negative elements $U^{(l)} \in \mathbb{R}_+^{m \times K}$ and $V^{(l)} \in \mathbb{R}_+^{K \times d^{(l)}}$, $K \ll d^{(l)}$, which factorization is formulated as $X^{(l)} \approx U^{(l)} V^{(l)}$. $U^{(l)}$ represents the class indicators, indicating the final clustering result. $V^{(l)}$ is termed the basis matrix. $K$ denotes the desired reduced dimension. The fundamental multi-view based on NMF function tries to minimize the joint problem over $U^{(l)}, V^{(l)}$:

$$\sum_{l}^{n_v} \|X^{(l)} - U^{(l)} V^{(l)}\|_F^2 + \Omega, \quad s.t.\, U^{(l)}, V^{(l)} \geq 0 \quad (1)$$

where $\Omega$ represents different kinds of penalty terms. $\|\cdot\|_F$ is the Frobenius norm of the matrix.

Table 1: Summary of the Notations

| Notations | Description |
|---|---|
| $m$ | Total number of items |
| $n_v$ | Total number of views |
| $d^{(l)}$ | Feature numbers in the $l$-th view |
| $X^{(l)}$ | Data matrix for the $l$-th view |
| $U^{(l)}$ | Class indicator matrix |
| $V^{(l)}$ | The basis matrix for the $l$-th view |
| $S^{(l)}$ | Semantic similarity matrix |
| $L^{(l)}$ | Laplacian matrix |
| $M^{(l)}$ | Sampling matrix for the $l$-th view |
| $\eta$ | Reduction factor |
| $\alpha, \beta$ | Parameters in objective function |

### Loss Function

In real-world application, feature space is much larger than item numbers(i.e. $d \gg m$). However, the cost of a single iteration depends linearly on dimension number $d$ and increases the computation time and memory requirements also degenerate performance of algorithms. We propose two feature sampling schemes which improve the efficiency of the update process for $V^{(l)}$. Our algorithm can be written as,

$$\begin{aligned} &\min_{U,V} \sum_{l}^{n_v} \|(X^{(l)} - U^{(l)} V^{(l)}) M^{(l)}\|_F^2 \\ &s.t. \quad {U^{(l)}}^T U^{(l)} = I, \quad U^{(l)}, V^{(l)} \geq 0 \end{aligned} \quad (2)$$

We enforce the orthogonal constraint on $U^{(l)}$ to guarantee the uniqueness of the solution. $M^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l)}}$ is a diagonal matrix with the $i$-th diagonal element $M_{i,i}^{(l)} \in \{0, 1\}$, which selects a subset of features in $X^{(l)}$ and $V^{(l)}$. Now we propose two strategies in our sampling process:

- **Top Sampling** The main intuition is to discard those features unobserved in most items (e.g, noise). We rank the columns in $X^{(l)}$ in descending order by using $l_2$-norm,

which is indicated by $\|X_{:,j}^{(l)}\|_2, j = 1, \cdots, d^{(l)}$. Then select top-$p$ columns as our sampled features. We should note that different view matrices have different feature numbers. Thus $p^{(l)} = \frac{d^{(l)}}{\eta} = \sum_{i=1}^{d^{(l)}} M_{i,i}$, where $\eta$ is a reduction factor.

- **Random Sampling** According to the top sampling method, sampled features are always the same at per iteration, which incurs the loss of important information, for instance, special features in special items. Therefore, we adopt random scheme in feature sampling process. Sampling number is $p^{(l)} = \frac{d^{(l)}}{\eta} = \sum_{i=1}^{d^{(l)}} M_{i,i} < d^{(l)}$. This strategy is more flexible than top sampling method.

### Unsupervised Semantic Regularization

**Intra-semantic Graph Regularizer** Our class indicator matrices $U^{(l)} = [u_1^{(l)}, u_2^{(l)}, \cdots, u_m^{(l)}]^T, u_i^{(l)} \in \mathbb{R}^{1\times K}, i = 1, 2, \cdots, m$. Then, our intra-semantic similarity graph regularizer can be formulated as,

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} S_{i,j}^{(l)}\|u_i^{(l)} - u_j^{(l)}\|_2^2 \\
&= \frac{1}{2}(2\sum_{i=1}^{m} d_i^{(l)} u_i^{(l)} {u_i^{(l)}}^T - 2\sum_{i=1}^{m}\sum_{j=1}^{m} S_{i,j}^{(l)} u_i^{(l)} {u_j^{(l)}}^T) \\
&= \mathrm{tr}({U^{(l)}}^T D^{(l)} U^{(l)}) - \mathrm{tr}({U^{(l)}}^T S^{(l)} U^{(l)}) \\
&= \mathrm{tr}({U^{(l)}}^T L^{(l)} U^{(l)})
\end{aligned} \tag{3}$$

where $L$ is a symmetric graph Laplacian matrix constructed from similarity matrix $S$, $D$ is a diagonal matrix with its elements defined as $d_i = D_{i,i} = \sum_{j=1}^{m} S_{i,j}$. Here, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ is used in above equation. The key task in our work is how to define unsupervised semantic similarity matrix $S \in \mathbb{R}^{m\times m}$, which is defined by two strategies,

- **Semantic Cosine based (SC)** The true underlying semantic information would assign corresponding items across different views into the same latent topic distribution, which can be indicated by $T(x_i^{(1)}) = T(x_i^{(2)}) = \cdots = T(x_i^{(n_v)}), i = 1, 2, \cdots, m$. Function $T(\cdot)$ returns the semantic topics vector, i.e., $T(x_i^{(l)}) = [f_1, f_2, \cdots, f_t] \in \mathbb{R}^{1\times t}$, $t$ is a topic number. We calculate pairwise similarity via the cosine kernel:

$$S_{i,j}^{(l)} = \begin{cases} \frac{<T(x_i^{(l)}), T(x_j^{(l)})>}{\|T(x_i^{(l)})\|\|T(x_j^{(l)})\|} & i \neq j \\ 1 & i = j \end{cases}$$

- **Gaussian kernel based (GK)** We construct matrix $S$ through Gaussian kernel: $S_{i,j}^{(l)} = \exp(-\frac{\frac{1}{n_v}\sum_{l=1}^{n_v}\|x_i^{(l)} - x_j^{(l)}\|_2^2}{\sigma^2})$ where $\sigma$ is the controlling parameter selected by cross-validation.

**Inter-semantic Diversity Term** In real web content applications, different views may be generated heterogeneously and may vary drastically in quality. To implement the inter-semantic diversity information of multi-view clustering, an intuitive method is to regularize the class indicator matrices $U^{(l)}$ of the different views, which are enforced to be independent to each other. We let $\mathcal{X} = \{\mathcal{X}_1, \mathcal{X}_2, \cdots, \mathcal{X}_{n_v}\}$ be our space of different views, and each $i$-th view is drawn from $\mathcal{X}_i \in \mathcal{X}$ space. Our purpose is to maximize the diversity and minimize the dependence between latent representations in different spaces. i.e. $\mathcal{X}_i \times \mathcal{X}_j$. Therefore, we utilize the *Hilbert Schmidt Independence Criterion* (HSIC) to deal with the dependence, which computes the sum of the squared singular values of the cross-covariance operator over $\mathcal{X}_i \times \mathcal{X}_j = \{(u_1^{(i)}, u_1^{(j)}), (u_2^{(i)}, u_2^{(j)}), \cdots, (u_m^{(i)}, u_m^{(j)})\}$ in Hilbert space and demonstrates fast exponential convergence. Formally, the HSIC (Arthur et al. 2005) is defined as

$$HSIC(U^{(l)}, U^{(s)}) = (m-1)^{-2}\,\mathrm{tr}(K^{(l)} Y K^{(s)} Y) \tag{4}$$

where $K^{(l)}, K^{(s)} \in \mathbb{R}_+^{m\times m}$ are Gram matrices of kernel functions. We employ the inner product kernel here, which means $K^{(l)} = U^{(l)}{U^{(l)}}^T$. $Y = I - \frac{1}{m}ee^T$, where $I$ is an identity matrix and $e$ is an all-one column vector. We use the HSIC as a penalty term in our objective function to ensure that representations in different views provide inter-semantic diversity information.

### Objective Function

According to the above introduction, the overall objection function is rewritten as follows:

$$\begin{aligned}
&\min_{U,V}\sum_{l}^{n_v}\|(X^{(l)} - U^{(l)}V^{(l)})M^{(l)}\|_F + \\
&\alpha\sum_{l\neq s} HSIC(U^{(l)}, U^{(s)}) + \beta\sum_{l}^{n_v}\mathrm{tr}({U^{(l)}}^T L^{(l)} U^{(l)}))
\end{aligned} \tag{5}$$

where $\alpha, \beta$ are the tradeoff parameters to control the weight between unsupervised semantic information and the proposed error loss function. For simplicity, we ignore the constant factor in the HSIC function, and Eq. 5 can be rewritten as

$$\begin{aligned}
&J_1 = \min_{U,V}\sum_{l}^{n_v}\|(X^{(l)} - U^{(l)}V^{(l)})M^{(l)}\|_F + \\
&\alpha\sum_{l\neq s}\mathrm{tr}(K^{(l)} Y K^{(s)} Y) + \beta\sum_{l}^{n_v}\mathrm{tr}({U^{(l)}}^T L^{(l)} U^{(l)})) \\
&s.t.\quad {U^{(l)}}^T U^{(l)} = I, \quad U^{(l)}, V^{(l)} \geq 0
\end{aligned} \tag{6}$$

By encouraging diversity of views, the algorithm finally learns a very different cluster class indicator matrix $U^{(l)}$ for each view $l$. The final aggregated representation $U$ can be obtained by combining all $U^{(l)}$: $U = [U^{(1)}, U^{(2)}, \cdots, U^{(n_v)}] \in \mathbb{R}^{m\times(n_v K)}$.

## 3 Optimization Algorithm

### Algorithms & Optimization

The objective function in Eq.6 is separately convex w.r.t each of $U^{(l)}$ and $V^{(l)}$. We handle this problem via alternating optimization. i.e., updating one variable while fixing

**Algorithm 1:** $FSUSC$ algorithm

**Input:** Data matrices $X^{(l)}$; Parameters $K, \alpha, \beta, \gamma, \eta$; Semantic similarity matrices $S^{(l)}$
**Output:** $U^{(l)}$, $V^{(l)}$;

```
Initialize U^(l) and V^(l) using the k-means;
Topic modeling for X^(l) and calculate S under different strategies(SC, GK);
repeat
    for l to n_v do
        Draw a feature sampling matrix: M^(l)(Top or Random);
        Update V^(l) by Eq. 8;
        Update U^(l) by Eq. 15;
    end
    return U^(l) and V^(l)
until Convergence;
```

other variables until convergence. Therefore, we propose an alternating optimization algorithm that guarantees each subproblem converges to the local minima under non-negative condition. The $FSUSC$ algorithm is summarized in Algorithm 1.

**Fixing $U^{(l)}$, then minimize $V^{(l)}$.** Let $\varphi$ and $\psi^{(l)}$ be the Lagrange matrices for constraint $U^{(l)} \geq 0$ and $V^{(l)} \geq 0$, respectively. For notional convenience, we let $\widetilde{X}^{(l)}, \widetilde{V}^{(l)}$ represent $X^{(l)}M^{(l)}, V^{(l)}M^{(l)}$. Then, the derivative of $J_1$ with respect to $V^{(l)}$ is:

$$\frac{\partial J_1}{\partial V^{(l)}} = (-2{U^{(l)}}^T \widetilde{X}^{(l)} + 2{U^{(l)}}^T U^{(l)} \widetilde{V}^{(l)}) + \psi^{(l)} \quad (7)$$

Using the Karush-Kuhn-Tucker(KKT) conditions that $\psi_{i,j}^{(l)} V_{i,j}^{(l)} = 0$, we have:

$$V_{i,j}^{(l)} \leftarrow V_{i,j}^{(l)} \frac{({U^{(l)}}^T X^{(l)} M^{(l)})_{i,j}}{({U^{(l)}}^T U^{(l)} V^{(l)} M^{(l)})_{i,j}} \quad (8)$$

**Fixing $V^{(l)}$, then minimize $U^{(l)}$.** Now, we analyze the stationary point $U^{(l)}$ in the second subproblem. Let $\gamma$ be the balance parameter for orthogonal constraint. So our minimization subproblem $J_2(U^{(l)})$ can be written as:

$$\begin{aligned} &\sum_{l}^{n_v} \{\mathrm{tr}((\widetilde{X}^{(l)})^T \widetilde{X}^{(l)}) - 2\,\mathrm{tr}((\widetilde{X}^{(l)})^T U^{(l)} \widetilde{V}^{(l)}) + \\ &\mathrm{tr}((\widetilde{V}^{(l)})^T {U^{(l)}}^T U^{(l)} \widetilde{V}^{(l)}) + \beta\,\mathrm{tr}({U^{(l)}}^T L^{(l)} U^{(l)})) \\ &+ \mathrm{tr}(\varphi U^{(l)}) + \gamma\,\mathrm{tr}({U^{(l)}}^T U^{(l)} - I)\} \\ &+ \alpha \sum_{l \neq s}^{n_v} \mathrm{tr}(U^{(l)} {U^{(l)}}^T Y K^{(s)} Y) \end{aligned} \quad (9)$$

Then the derivative of $J_2$ with respect to $U^{(l)}$ is:

$$\begin{aligned} &-2\widetilde{X}^{(l)}(\widetilde{V}^{(l)})^T + 2U^{(l)}\widetilde{V}^{(l)}(\widetilde{V}^{(l)})^T + 2\beta LU \\ &+ 2\gamma U + \varphi + 2\alpha \sum_{l \neq s}^{n_v} Y K^{(s)} Y U^{(l)} \end{aligned} \quad (10)$$

However, $Y$ in above equation contains negative values, we let $Y = Y^+ - Y^-$, which is separated to two nonnegative parts. And $Y_{pq}^+ = \frac{\|Y_{pq}\| + Y_{pq}}{2}, Y_{pq}^- = \frac{\|Y_{pq}\| - Y_{pq}}{2}$. Following the KKT rule, we have the following equation if we set the Eq. 10 to be 0 and we can easily get the update rule for $U^{(l)}$,

$$\begin{aligned} &(\widetilde{X}^{(l)}(\widetilde{V}^{(l)})^T - U^{(l)}\widetilde{V}^{(l)}(\widetilde{V}^{(l)})^T - \beta LU - \gamma U \\ &- \alpha(\sum_{l \neq s}^{n_v} Y^+ K^{(s)} Y^+ U^{(l)} + \sum_{l \neq s}^{n_v} Y^- K^{(s)} Y^- U^{(l)}) \\ &+ \alpha(\sum_{l \neq s}^{n_v} Y^+ K^{(s)} Y^- U^{(l)} + \sum_{l \neq s}^{n_v} Y^- K^{(s)} Y^+ U^{(l)}) \\ &)_{ip} U_{ip}^{(l)} = 0 \end{aligned} \quad (11)$$

Now we prove that Eq.9 keeps non-increasing when updating the stationary point $U^{(l)}$ from Eq. 11 .

**Definition 1** *(Lee and Seung 2001) $F(U, U')$ is an auxiliary function of $J_2$ if the conditions $F(U, U') \geq J_2$ and $F(U, U) = J_2$. If $F$ is an auxiliary function for $J_1$ then $J_1$ is non-increasing under the update*

$$U^{(t+1)} = \arg\min_{U} F(U, U^t) \quad (12)$$

**Lemma 1** *(Ding, Li, and Jordan 2010) For any matrices $A \in \mathbb{R}_+^{n \times n}, B \in \mathbb{R}_+^{r \times r}, Q \in \mathbb{R}_+^{n \times r}, Q' \in \mathbb{R}_+^{n \times r}$, with $A$ and $B$ symmetric, the following inequality holds:*

$$Tr(Q^T A Q B) \leq \sum_{i=1}^{n} \sum_{p=1}^{r} \frac{(AQ'_{ip}B)Q_{ip}^2}{Q'_{ip}} \quad (13)$$

First, we should denote some brief notations $Y_1 = \sum_{l \neq s}^{n_v} Y^+ K^{(s)} Y^-, Y_2 = \sum_{l \neq s}^{n_v} Y^- K^{(s)} Y^+, Y_3 = \sum_{l \neq s}^{n_v} Y^- K^{(s)} Y^-, Y_4 = \sum_{l \neq s}^{n_v} Y^+ K^{(s)} Y^+$ and we use $U$ to represent $U^{(l)}$ in the following equation. Second, we ignore the irrelevant terms in Eq.9 and an appropriate auxiliary function $F(U, U')$ of $J_2$ is as the following:

$$\begin{aligned} &-\sum_{ip} 2(X^{(l)} M^{(l)} (V^{(l)})^T)_{ip} U'_{ip} (1 + \log \frac{U_{ip}}{U'_{ip}}) \\ &+ \sum_{ip} [U'(V^{(l)} M^{(l)})(V^{(l)} M^{(l)})^T]_{ip} \frac{U_{ip}^2}{U'_{ip}} \\ &- \beta \sum_{ijpq} S_{ip} U'_{ip} U'_{jq} (1 + \log \frac{U_{ip} U_{jq}}{U'_{ip} U'_{jq}}) \\ &+ \beta \sum_{ip} (DU')_{ip} \frac{U_{ip}^2}{U'_{ip}} + \gamma \sum_{ip} (U')_{ip} \frac{U_{ip}^2}{U'_{ip}} \\ &- \alpha \sum_{ijpq} ((Y_1 + Y_2)_{ip} U'_{ip} U'_{jq} (1 + \log \frac{U_{ip} U_{jq}}{U'_{ip} U'_{jq}}) \\ &+ \alpha \sum_{ip} ((Y_3 + Y_4) U')_{ip} \frac{U_{ip}^2}{U'_{ip}} \end{aligned} \quad (14)$$

All negative terms in $F(U, U')$ are produced by the following inequality: $z \geq 1 + \log z, \forall z \geq 0$ . Now we should find

Table 2: Description of datasets

| Dataset | #items size | #view | #clusters |
|---|---|---|---|
| Doucom | 31297 | 4 | 39 |
| Last.fm | 9694 | 3 | 21 |
| Yelp | 2624 | 3 | 7 |
| 3-Sources | 169 | 3 | 6 |

the minimum of $F(U, U^{'})$, we set $\frac{\partial F(U,U^{'})}{\partial U_{ip}}$ to zero, we can get the stationary point of $U_{ip}$ and the updating rule is:

$$U_{ip}\sqrt{\frac{(\lambda_l \widetilde{X}^{(l)}(V^{(l)})^T + \beta SU + \alpha U(Y_1+Y_2))_{ip}}{(\lambda_l U\widetilde{V}^{(l)}(\widetilde{V}^{(l)})^T + \beta DU + \gamma U + \alpha U(Y_3+Y_4))_{ip}}} \tag{15}$$

Then, we take the second derivative with respect to $U$, we get a positive semidefinite Hessian matrix:

$$\begin{aligned}\frac{\partial^2 F(U,U^{'})}{\partial U_{ip}\partial U_{jq}} = \{&\frac{2\lambda_l(\widetilde{X}^{(l)}(V^{(l)})^T)_{ip}U^{'}_{ip} + 2\beta S_{ip}(U^{'}_{ip})^2}{U^2_{ip}} \\ &+ \frac{2\alpha(U^{'}(Y_1+Y_2))_{ip}U^{'}_{ip}}{U^2_{ip}} + \frac{2[\lambda_l U^{'}(\widetilde{V}^{(l)})(\widetilde{V}^{(l)})^T]_{ip}U^2_{ip}}{(U^{'}_{ip})^3} \\ &+ \frac{2[\beta(DU^{'})_{ip} + \gamma(U^{'})_{ip} + \alpha(U^{'}(Y_3+Y_4))_{ip}]U^2_{ip}}{(U^{'}_{ip})^3}\}\zeta_{ij}\zeta_{jq}\end{aligned} \tag{16}$$

Thus $F(U, U^{'})$ is a convex function of $U$, and under the updating rule Eq. 15 the objective function values of $J_2(U^{(l)})$ in Eq. 9 will be non-increasing.

### Complexity Analysis

There are two subproblems in $FSUSC$ algorithm: optimizing $V^{(l)}$ and optimizing $U^{(l)}$. It can be shown that cost for Multi-view NMF's update rules in each iteration is $O(n_v K m d^{(l)})$. But the computation cost for updating $V^{(l)}$ is $O(n_v K m p^{(l)})$ after applying the feature sampling. For class indicator matrix $U^{(l)}$, the largest cost is the first term of the denominator, whose time complexity is $O(n_v K^2 m p^{(l)})$. $n_v$ denotes the number of views. The total time complexity is $O(n_v K m p^{(l)}) + O(n_v K^2 m p^{(l)}) \approx O(n_v K^2 m p^{(l)})$, which is a linear to the size of dataset. In a real application, the number of items and clusters are much smaller than the size of features(e.g. $K, m \ll d^{(l)}$ ). Our method can be suitable for large-scale data if we utilize feature sampling in the iteration process.

## 4 Experiments

In this section, we conduct experiments to evaluate the effectiveness of $FSUSC$.

### Datasets and Settings

Table 2 summarizes the characteristics of those real web multi-view datasets and all descriptions of the datasets are as follows.

1)**Doucom.** This large-scale dataset is crawled from a famous web community, called $Douban$[2], we collect four views for this dataset, including 31297 summaries, 2,995,406 comments, 608,158 reviews and 461,358 users. After data preprocessing, we have 50,992; 46,706; 32,935; 232,531 token features for view "Summary", "Long review", "Short commentary", "Users" respectively. $Doucom$ lists 39 movie types such as "Action", "Love", "Bloopers", etc. In fact, for some items tagged with multiple types, we retain one type to annotate single item.

2)**Last.fm.** This dataset consists of 9,694 items(artists), which contains three views such as description of each item, user comments and users. Each item is annotated with one of the 21 music genres. All the relevant textual information can be achieved by API of Last.fm[3].

3)**Yelp.** This dataset is a subset of the Yelp Challenge Dataset (YDC)[4], which includes 11,537 items (businesses) in total. We randomly sample the equivalent amount of items from the every category. So final Yelp dataset has three views (eg. businesses' names view, comment words view and user view) and consists of 2,624 items from 7 categories.

4) **3-Sources.**[5] This text dataset was collected from three well-known online news sources: BBC, Reuters and The Guardian. In total it consists of 416 distinct news manually categorized into 6 topical labels. Among them, there are 169 stories reported in all three sources which are used as three views in our experiments.

To evaluate the performance of the proposed method, we compare our method with the following algorithms.**CoRe.** (Kumar, Rai, and Daumé 2011) proposed the objective functions to co-regularize the eigenvectors of all views' Laplacian matrices. **MulitNMF** (Liu et al. 2013) developed a solution on consensus-based regularization for NMF to group the multi-view data. **PcoNMF** (He et al. 2014) is a recent pair-wise co-regularization method for clustering the whole mapped data, which focus on the difference between the latent indicator matrix. **CMVNMF** (Zhang et al. 2015) proposed a novel small number of constraints on must-link sets and cannot-link sets based on the NMF framework. And we let $FSUSC$-**intra** be our baseline method that only includes intra-semantic similarity constraint and let $FSUSC$-**inter** be our another baseline method that only includes inter-semantic diversity constraint.

In this work, we have six major parameters: { $\sigma$, $\alpha$, $\beta$, $\gamma$, $\eta$, $K$}. We empirically set $\sigma = 0.01$ in kernel function. $\alpha, \beta, \gamma$ are the weight coefficients, which are set to $1, 2, 1$, respectively. $\eta$ is a reduction factor which controls the different size of feature subset, we set $\eta = 8$ in final results and compare the computation time with different value at per iteration(See Fig.1. and Fig.3.). Also we compare the time consumption of the proposed algorithm with necessary baseline models in Table 3. $K$ is a reduced dimension number that equals to cluster numbers which are described in each dataset. We extract topic vector as our unsupervised semantic information to build the similarity matrix. Each item is represented as a 20-dimensional topical feature using Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003). For all the

[2]https://developers.douban.com/wiki/?title=api_v2

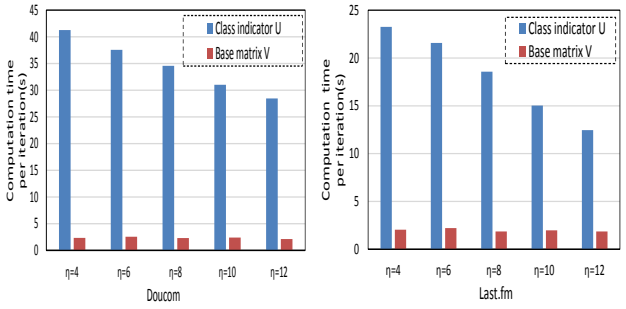
[3]http://www.last.fm/api

[4]http://www.yelp.com/dataset_challenge

[5]http://mlg.ucd.ie/datasets

Table 3: Average computation time for convergence of different methods(in seconds)

| | **Doucom** | **Last.fm** | **Yelp** | **3-Sources** |
|---|---|---|---|---|
| CoRe | 32,544 | 4,821 | 82 | 14 |
| MultiNMF | 40,451 | 7,768 | 137 | 36 |
| PcoNMF | 33,693 | 5,209 | 65 | 23 |
| CMVNMF | 30,981 | 5,670 | 146 | 60 |
| $FSUSC$ | 10,732 | 2,771 | 60 | 26 |




(a) (b)

Figure 1: Iteration time with different reduction factor

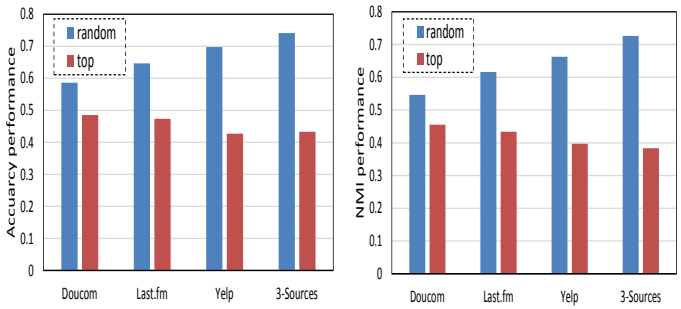



Figure 2: Sampling strategy with respect to clustering performance on each dataset

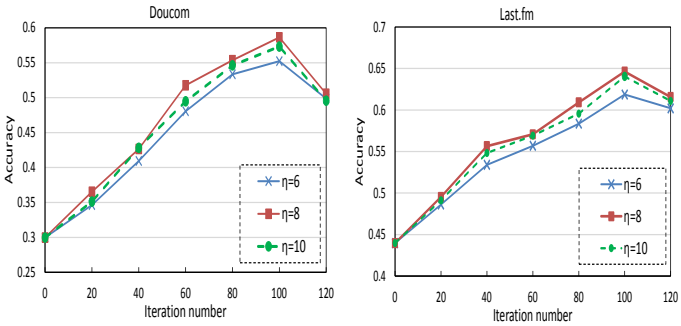



Figure 3: The best performance with various parameter $\eta$

used text datasets, we apply the TF-IDF transformation on all item-word frequency matrices.To evaluate the clustering performance, we use clustering accuracy(ACC) and normalized mutual information(NMI) (Du, Li, and Shen 2012) as our metrics. The larger ACC and NMI are, the better performance is. We put all view matrices together to form a huge one, then we run K-means 100 times and select the best clustering result to initialize our factor $U^{(l)}$ and $V^{(l)}$.

## Clustering Results

Comparison results of our two sampling strategies are shown in Fig.2.. It is obvious that the efficiency of random sampling strategy is much better than top sampling strategy. It should be noted that the feature subset is fixed per iteration in top sampling method, and this situation deeply influences the clustering performance, especially in some small datasets like **Yelp** and **3-Sources**. Stationary feature subset results in the information loss which is the only reasonable cause for performance sharply decreasing in **Yelp** and **3-Sources**. Our random sampling strategy is more flexible than top sampling in above aspect. Therefore, the random sampling strategy is used as our final method in all experiments. We compared two unsupervised similarity information extractors and supervised label similarity extractors in Table 5 and 6, we can also observe that combining the unsupervised semantic prior information with our loss function outperforms supervised label similarity measurements, especially in **Doucom**&**Last.fm**. Because supervised manual label information represents a matrix for all views of content. Further, label information is a limited source in large scale web datasets and degenerates clustering performance in our framework. Further, we should know that **Doucom** and **Last.fm** are sparser than other datasets because of their dimensional problem, and this further indicates that our unsupervised semantic regularization on a latent factor matrix is a better solution to sparseness problem. In fact, GK uses the whole feature space to calculate similarity, and an entire feature space suffers from sparseness problem. This is a strong reason that declines the performance of our $FSUSC$ framework.

In Table 4, we present results of all methods measured by ACC and NMI for each dataset. We observe that: 1)The $FSUSC$ framework usually achieves a better improvement in all web datasets. This may indicate that $FSUSC$ framework has the evident effect in each web dataset, especially more suitable for real large corpus in the web application. 2)Among the NMF based clustering methods with different similarity constraint, a framework with the intra-semantic similar constraint and inter-semantic diversity information (e.g. $FSUSC$) performs much better than the simple pairwise constraint, which validates that the algorithm based on our proposed unsupervised semantic framework might be a better way of capturing the difference in intrinsic connection between every two data points. 3) $FSUSC$-inter shows a promising result, especially in **Doucom**. That means our inter-semantic constraint could capture diversity information among different views in large datasets, which contains many noises and useless features.

## Parameter study

Reduction factor $\eta$ properly determines the feature numbers that we should use in updating process. Relative $\alpha$

Table 4: Performance on four real-world datasets(Both mean value and standard deviation are reported,best results are formated in bold, while second best result are underlined)

| | ACC(%) | | | | NMI(%) | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | **Doucom** | **Last.fm** | **Yelp** | **3-Sources** | **Doucom** | **Last.fm** | **Yelp** | **3-Sources** |
| CoRe | 40.5 (±2.9) | 48.7 (±2.4) | 58.8 (±2.7) | 46.3 (±0.6) | 35.6 (±3.1) | 46.6 (±3.4) | 53.2 (±3.3) | 40.6 (±0.2) |
| MultiNMF | 40.1 (±4.7) | 45.5 (±2.3) | 30.2 (±2.6) | 68.4 (±0.1) | 37.6 (±4.2) | 39.4 (±2.3) | 34.7 (±1.9) | 60.2 (±0.1) |
| PcoNMF | 46.3 (±3.6) | 51.8 (±2.5) | 63.6 (±4.6) | 69.3 (±1.8) | 44.8 (±3.7) | 47.6 (±2.1) | 61.7 (±3.2) | 68.2 (±3.6) |
| CMVNMF | 51.6 (±7.1) | 60.4 (±3.8) | 64.3 (±3.8) | 70.9 (±5.7) | 46.6 (±6.1) | 57.2 (±1.8) | 62.4 (±2.7) | 68.4 (±5.5) |
| $FSUSC$-intra | 43.2 (±4.1) | 48.3 (±2.2) | 65.2 (±3.5) | 71.8 (±0.5) | 40.2 (±3.8) | 45.4 (±3.1) | 63.2 (±1.2) | 68.8 (±2.6) |
| $FSUSC$-inter | 52.5 (±5.6) | 58.3 (±2.1) | 67.2 (±4.4) | 73.2 (±2.2) | 49.5 (±4.3) | 56.1 (±1.3) | 64.6 (±4.2) | 71.9 (±3.5) |
| $FSUSC$ | **58.6** (±3.3) | **64.6** (±2.1) | **70.7** (±3.2) | **74.1** (±1.3) | **54.6** (±2.6) | **61.7** (±1.4) | **67.2** (±0.9) | **72.6** (±2.3) |

Table 5: ACC with different semantic constraints

| Data | Doucom | Last.fm | Yelp | 3-Sources |
|---|---|---|---|---|
| Supervised | 45.6 | 54.8 | 63.3 | 69.6 |
| $FSUSC$ +GK | 50.4 | 57.8 | 65.7 | 72.8 |
| $FSUSC$ +SC | **58.6** | **64.6** | **70.7** | **74.1** |

Table 6: NMI with different semantic constraints

| Data | Doucom | Last.fm | Yelp | 3-Sources |
|---|---|---|---|---|
| Supervised | 43.7 | 52.1 | 62.4 | 67.9 |
| $FSUSC$ +GK | 49.6 | 56.4 | 63.9 | 71.3 |
| $FSUSC$ +SC | **54.6** | **61.7** | **67.2** | **72.6** |

and $\beta$ determine the weights of the intra-semantic similarity graph regularizer and inter-semantic diversity term, respectively. Figure 3 shows the performance of different reduction factor on two large real-world datasets and Fig.4. evaluates our constraint parameter while holding $\gamma = 1, \eta = 8$ for all views on the two largest datasets. We also studied whether we could improve the clustering by tuning the parameter $\gamma$. However, the performance is not improved much. In Fig.3., our iteration number is set to be 120, our algorithm will converge during 100 iterations. And we can observe the performance curve in **Doucom** shows more steepness because **Doucom** is much larger than **Last.fm**. In Fig.4, $FSUSC$ performs the sensitive change, which illustrates that $FSUSC$ is more suitable for the large and sparse corpus. In addition, $FSUSC$ performs best when $\alpha, \beta$ locates in 1 or 2 when varying $\gamma$ for all views. We also studied the parameter on other small datasets like **Yelp** and **3-Sources**, and all results indicate that performance is the best when $\alpha, \beta$ locates around 1 or 2. This suggests that the parameter

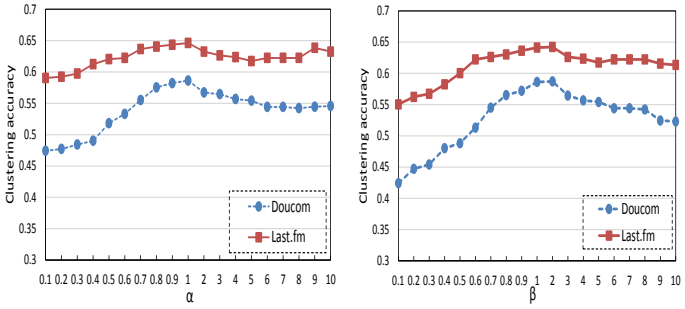



Figure 4: Evaluating parameters $\alpha$ and $\beta$ with respect to clustering accuracy on two large datasets

$\alpha, \beta$ can be set to 1 and 2, respectively.

## 5 Conclusion

We have proposed $FSUSC$ framework, a featured sampling based nonnegative matrix factorization algorithm which is combined with unsupervised semantic constraint, which can handle the real-world multiple view datasets with a large number of features. Also, we have developed an iterative optimization algorithm to make iteration faster and accelerate convergence. Extensive experiments have demonstrated that the proposed method is effective. In the future, we will study how to model any other features together generated by comment users such as the list of user preference and investigate how to improve the algorithm efficiency when dealing with the huge items and features both.

## Acknowledgments

This work is supported by National Key Research and Development Program of China under grant No. 2018YFB1003302, and National Natural Science Foundation of China under grant No.61472241. Linpeng Huang is the corresponding author of this paper.

## References


Arthur, G.; Olivier, B.; Alex, S.; and Bernhard, S. 2005. Measuring statistical dependence with hilbert-schmidt norms. *In International conference on algorithmic learning theory, Springer* 63–77.

Bickel, S., and Scheffe, T. 2004. Multi-view clustering. *IEEE International Conference on Data Mining,*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *JMLR,* 3:993–1022.

Cai, X.; Nie, F.; and Huang, H. 2013. Multi-view k-means clustering on big data. *International joint conference on Artificial Intelligence.* 2598–2604.

Chaudhuri, K.; Kakade, S. M.; Livescu, K.; and Sridharan, K. 2009. Multi-view clustering via canonical correlation analysis. *In ICML,New York, NY, USA*.

Cheng, W.; Zhang, X.; Guo, Z.; and Wu, Y. 2013. Flexible and robust co-regularized multi-domain graph clustering. *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining-KDD'13.* 1:320–328.

Ding, C. H.; Li, T.; and Jordan, M. I. 2010. Convex and semi-nonnegative matrix factorizations. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 32(1):45–55.

Du, L.; Li, X.; and Shen, Y. 2012. Robust nonnegative matrix factorization via half-quadratic minimization. *In Proceedings of the 12th International Conference on Data Mining,* 201–210.

Gong, X.; Wang, F.; and Huang, L. 2017. Weighted nmf-based multiple sparse views clustering for web items. *In PAKDD'17.* 416–428.

Greene, D., and Cunningham, P. 2009. A matrix factorization approach for integrating multiple data views. *European Conference on Machine Learning and Knowledge Discovery in Databases.* 423–438.

He, X.; Kan, M.; Xie, P.; and Chen, X. 2014. Comment-based multi-view clustering of web 2.0 items. *In International Conference on World Wide Web,* 771–782.

Hoyer, P. O. 2004. Nonnegative matrix factorization with sparseness constraints. *J. Machine Learning Research,* 5:1457–1469.

J. Sun, J. B., and Kranzler, H. R. 2014. Multi-view singular value decomposition for disease subtyping and genetic associations. *BMC Genet.* 15(73).

Kumar, A., and Daumé, H. 2011. A co-training approach for multi-view spectral clustering. *In ICML'11,* 393–400.

Kumar, A.; Rai, P.; and Daumé, H. 2011. Co-regularized multi-view spectral clustering. *In Proc. of NIPS'11,* 1413–1421.

Lee, D. D., and Seung, H. S. 2001. Algorithms for non-negative matrix factorization. *In Advances in neural information processing systems,* 13:252–260.

Liu, J.; Wang, C.; Gao, J.; and Han, J. 2013. Multi-view clustering via joint nonegative matrix factorization. *In Proc. of SDM'13,* 252–260.

Long, B.; Yu, P. S.; and Zhang, Z. 2008. A general model for multiple view unsupervised learning . *SIAM International Conference on Data Mining.* 822–833.

Mohammadiha, N., and Leijon, A. 2009. Nonnegative matrix factorization using projected gradient algorithms with sparseness constraints. *Proc. IEEE Int'l Symp. Signal Processing and Information Technology,* 418–423.

Sun, J.; Lu, J.; Xu, T.; and Bi, J. 2015. Multi-view sparse co-clustering via proximal alternating. *In Proceedings of the 32th International Conference on Machine Learning, Lille, France,* 37.

Wang, Z.; Yang, Y.; Chang, S.; Li, J.; Fong, S.; and Huang, T. S. 2015. A joint optimization framework of sparse coding and discriminative clustering. *In Proceedings of the 24th International Joint Conference on Artificial Intelligence.* 3932–3938.

Yang, Y.; Wang, Z.; Yang, J.; Wang, J.; Chang, S.; and Huang, T. S. 2014. Data clustering by laplacian regularized l1-graph. *In Proceedings of the 28th AAAI Conference on Artificial Intelligence.*

Zhai, D.; Chang, H.; Zhen, Y.; Liu, X.; Chen, X.; and Gao, W. 2013. Parametric local multimodal hashing for cross-view similarity search. *In Proceedings of the 23rd International Joint Conference on Artificial Intelligence.* 2754–2760.

Zhang, X.; Zong, L.; Liu, X.; and Yu, H. 2015. Constrained nmf-based multi-view clustering on unmapped data. *In AAAI'15,* 3174–3180.

Zheng, M.; Bu, J.; Chen, C.; Wang, C.; Zhang, L.; Qiu, G.; and Cai, D. 2011. Graph regularized sparse coding for image representation. *Image Processing, IEEE Transactions on,* 20(5):1327–1336.

Zhu, X.; Wu, X.; Ding, W.; and Zhang, S. 2013. Feature selection by joint graph sparse coding. *In SDM'13.*