HINMF: A Matrix Factorization Method for Clustering in Heterogeneous Information Networks

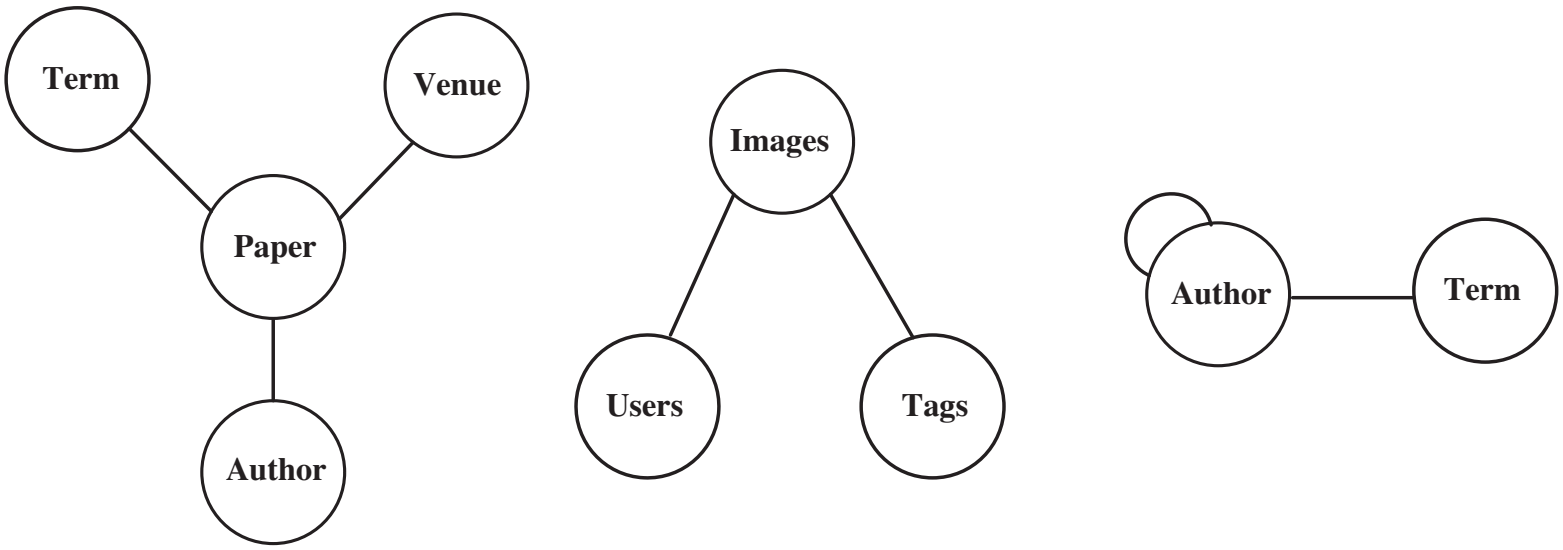
Jialu Liu Jiawei Han

University of Illinois at Urbana-Champaign

August 5, 2013

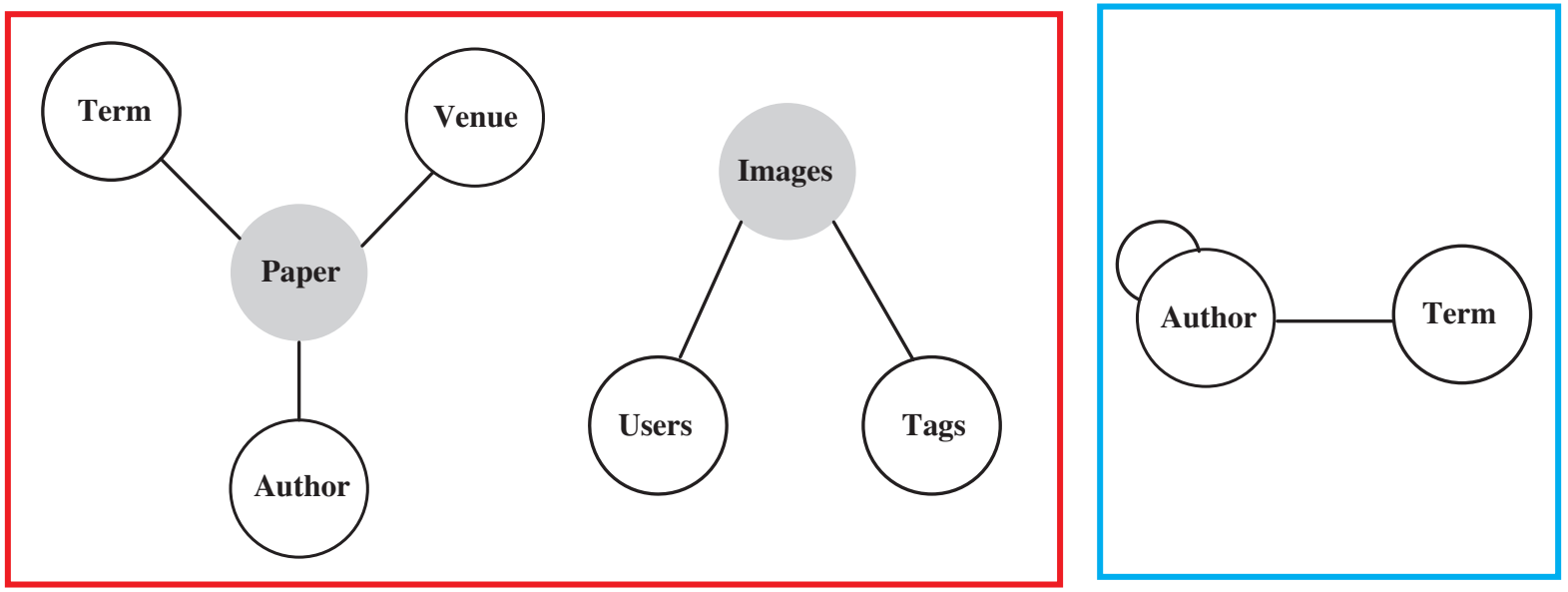
Outline

Heterogeneous Information Networks



In heterogeneous information networks (HIN), multiple types of nodes are connected by multiple types of links.

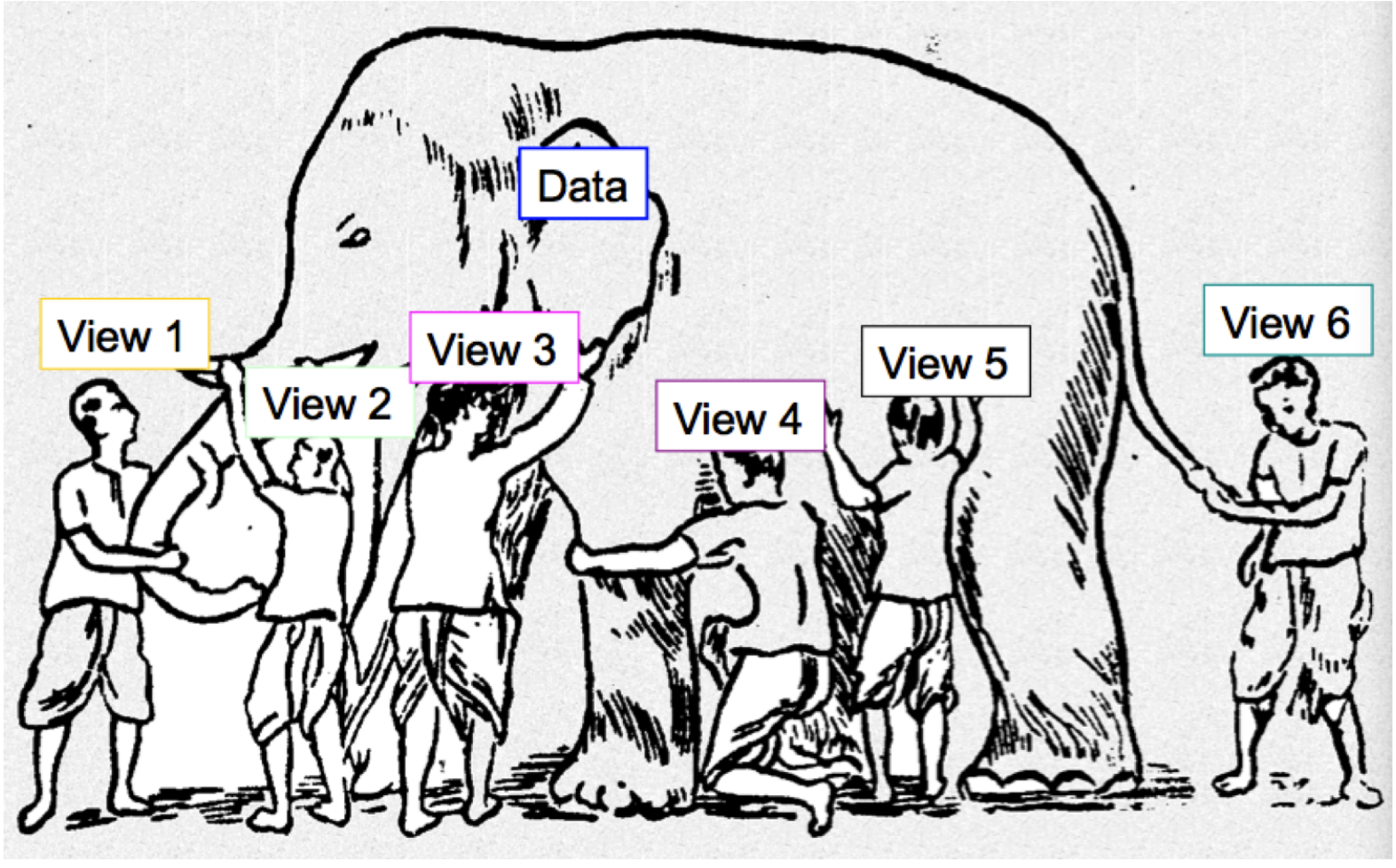
Star Schema



Grey: Center type, White: Attribute type

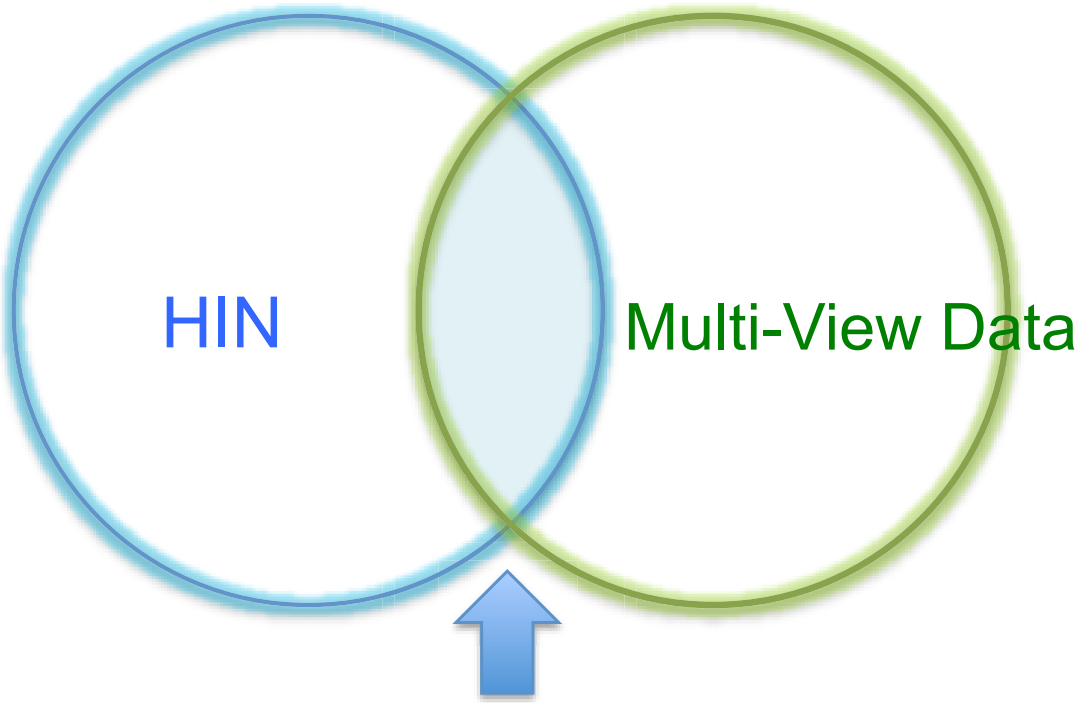
Multi-View Learning

Many datasets in real world are naturally comprised of different representations or *views*.



Connection between HIN and Multi-View data

HIN following star schema can be viewed as a kind of multi-view relational data. Attribute types provide “views” for the center type.



Common Motivation

Observing that multiple subnetworks/representations often provide *compatible* and *complementary* information, it becomes natural for one to integrate them together to obtain better performance rather than relying on a single homogenenous/bipartite network or view.

Outline

Nonnegative Matrix Factorization

Let $X = [X_{\cdot,1}, \ldots, X_{\cdot,N}] \in \mathbb{R}_+^{M \times N}$ denote the nonnegative data matrix where each column represents a data point and each row represents one attribute. NMF aims to find two non-negative matrix factors $U = [U_{i,k}] \in \mathbb{R}_+^{M \times K}$ and $V = [V_{j,k}] \in \mathbb{R}_+^{N \times K}$ whose product provides a good approximation to X:

$$X \approx UV^T \tag{1}$$

Here K denotes the desired reduced dimension, and to facilitate discussions, we call U the *basis matrix* and V the *coefficient matrix*.

Update Rule of NMF

One of the common reconstruction processes can be formulated as a *Frobenius norm* optimization problem, defined as:

$$\min_{U,V} ||X - UV^T||_F^2, \ s.t. \ U \geq 0, V \geq 0$$

Multiplicative update rules are executed iteratively to minimize the objective function as follows:

$$U_{i,k} \leftarrow U_{i,k} \frac{(XV)_{i,k}}{(UV^T V)_{i,k}}, \ V_{j,k} \leftarrow V_{j,k} \frac{(X^T U)_{j,k}}{(VU^T U)_{j,k}} \tag{2}$$

NMF for Clustering

Note that given the NMF formulation in Equation 1, for arbitrary invertible $K \times K$ matrix Q, we have

$$UV^T = (UQ^{-1})(QV^T) \tag{3}$$

There can be many possible solutions, and it is important to enforce constraints to ensure uniqueness of the factorization for clustering.

One of the common ways is to normalize basis matrix U after convergence of multiplicative updates if we use V for clustering:

$$U_{i,k} \leftarrow \frac{U_{i,k}}{\sqrt{\sum_i U_{i,k}^2}}, \; V_{j,k} \leftarrow V_{j,k} \sqrt{\sum_i U_{i,k}^2} \tag{4}$$

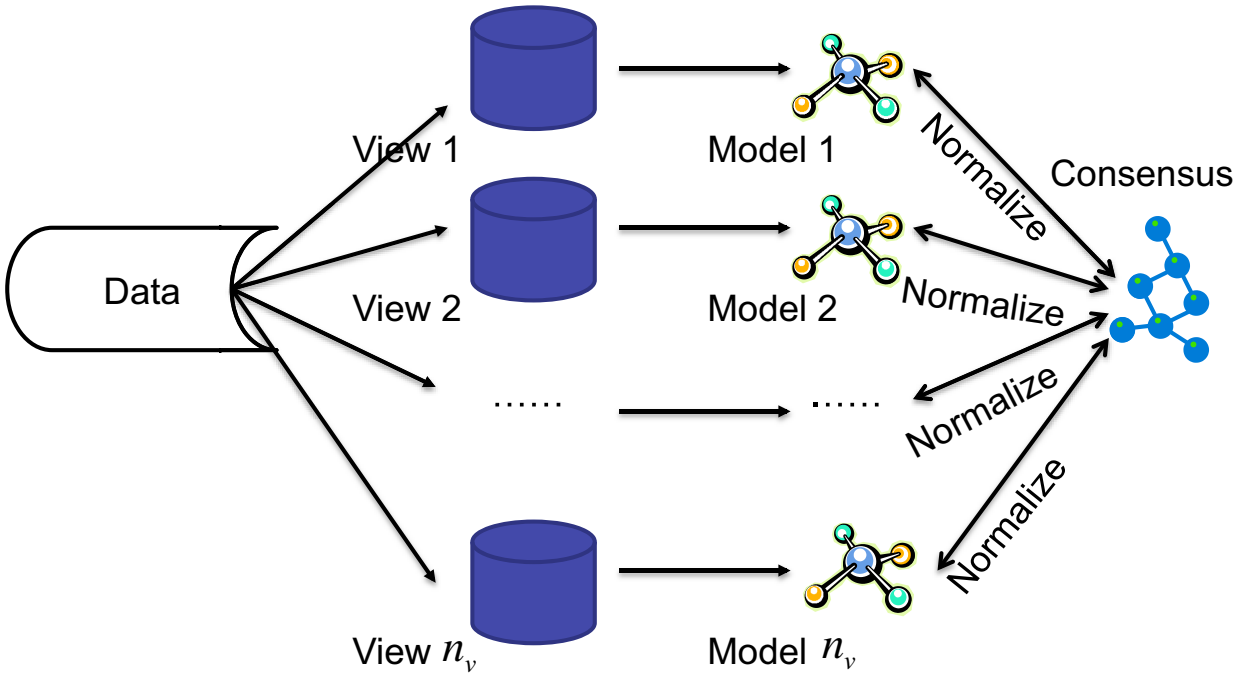
Outline

Multi-View Notations

Assume that we are now given n_v representations (i.e., views). Let $\{X^{(1)}, X^{(2)}, \ldots, X^{(n_v)}\}$ denote the data of all the views, where for each view $X^{(v)}$, we have factorizations that $X^{(v)} \approx U^{(v)}(V^{(v)})^T$.

Here for different views, we have the same number of data points but allow for different number of attributes, hence $V^{(v)}$s are of the same shape but $U^{(v)}$s can differ along the row dimension across multiple views.

Framework of MultiNMF



Models learnt from different views are requried to be softly regularized towards a consensus with proper normalization for clustering.

The Approach

Firstly, the disagreement between coefficient matrix $V^{(v)}$ and the consensus matrix V^* are incorporated into NMF:

$$\sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$
$$s.t.\ U^{(v)}, V^{(v)}, V^* \geq 0 \tag{5}$$

The Approach

Secondly, constraints on coefficient matrices $U^{(v)}$ in different views are added to make $V^{(v)}$s comparable and meaningful for clustering.

W.l.o.g., assume $||X^{(v)}||_1 = 1$, we then want to minimize:

$$\sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$
$$s.t.\ \forall 1 \leq k \leq K, ||U^{(v)}_{\cdot,k}||_1 = 1 \text{ and } U^{(v)}, V^{(v)}, V^* \geq 0 \qquad (6)$$

Why $||X||_1 = 1$ and $||U_{\cdot,k}||_1 = 1$?

Objective function:

$$\min_{U^{(v)}, V^{(v)}, V^*} \sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$

$$s.t.\ \forall 1 \le k \le K, ||U_{\cdot,k}^{(v)}||_1 = 1 \text{ and } U^{(v)}, V^{(v)}, V^* \ge 0$$

Given $||X||_1 = 1$ and $||U_{\cdot,k}||_1 = 1$,

$$||X||_1 = ||\sum_j X_j||_1 \approx \sum_{k=1}^{K} ||U_{\cdot,k} \sum_j V_{j,k}||_1 = \sum_{k=1}^{K} ||\sum_j V_{j,k}||_1 = ||V||_1$$

Therefore,

$$||V||_1 \approx 1$$

Objective Function

Previous:

$$\min_{U^{(v)},V^{(v)},V^*} \sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)} - V^*\|_F^2$$
$$s.t.\ \forall 1 \le k \le K, ||U^{(v)}_{\cdot,k}||_1 = 1 \text{ and } U^{(v)}, V^{(v)}, V^* \ge 0$$

Now:

$$\min_{U^{(v)},V^{(v)},V^*} \sum_{v=1}^{n_v} \|X^{(v)} - U^{(v)}(Q^{(v)})^{-1}Q^{(v)}(V^{(v)})^T\|_F^2 + \sum_{v=1}^{n_v} \lambda_v \|V^{(v)}Q^{(v)} - V^*\|_F^2$$
$$s.t.\ \forall 1 \le v \le n_v, U^{(v)} \ge 0, V^{(v)} \ge 0, V^* \ge 0$$

where

$$Q^{(v)} = Diag\left(\sum_{i=1}^{M} U^{(v)}_{i,1}, \sum_{i=1}^{M} U^{(v)}_{i,2}, \ldots, \sum_{i=1}^{M} U^{(v)}_{i,K}\right)$$

Iterative Update Rules

Fixing V^*, minimize over $U^{(v)}$ and $V^{(v)}$ until convergence:

$$U_{i,k} \leftarrow U_{i,k} \frac{(XV)_{i,k} + \lambda_v \sum_{j=1}^{N} V_{j,k} V^*_{j,k}}{(UV^T V)_{i,k} + \lambda_v \sum_{l=1}^{M} U_{l,k} \sum_{j=1}^{N} V^2_{j,k}}$$

$$U \leftarrow UQ^{-1}, \; V \leftarrow VQ$$

$$V_{j,k} \leftarrow V_{j,k} \frac{(X^T U)_{j,k} + \lambda_v V^*_{j,k}}{(VU^T U)_{j,k} + \lambda_v V_{j,k}}$$

Fixing $U^{(v)}$ and $V^{(v)}$, minimize over V^*:

$$V^* = \frac{\sum_{v=1}^{n_v} \lambda_v V^{(v)} Q^{(v)}}{\sum_{v=1}^{n_v} \lambda_v} \geq 0$$

Use V^* for Clustering

Once we obtain the consensus matrix V^*, the cluster label of data point j could be computed as $\arg\max_k V^*_{j,k}$.

Or we can simply use k-means directly on V^* where V^* is viewed as a latent representation of the original data points.

Outline

PLSA

Probabilistic Latent Semantic Analysis (PLSA) is a traditional topic modeling technique for document analysis. It models the $M \times N$ term-document co-occurrence matrix X (each entry X_{ij} is the number of occurrences of word w_i in document d_j) as being generated from a mixture model with K components:

$$P(w, d) = \sum_{k=1}^{K} P(w|k)P(d, k)$$

Relation to NMF

$$P(w,d) = \sum_{k=1}^{K} P(w|k)P(d,k)$$

$$X = (UQ^{-1})(QV^T)$$

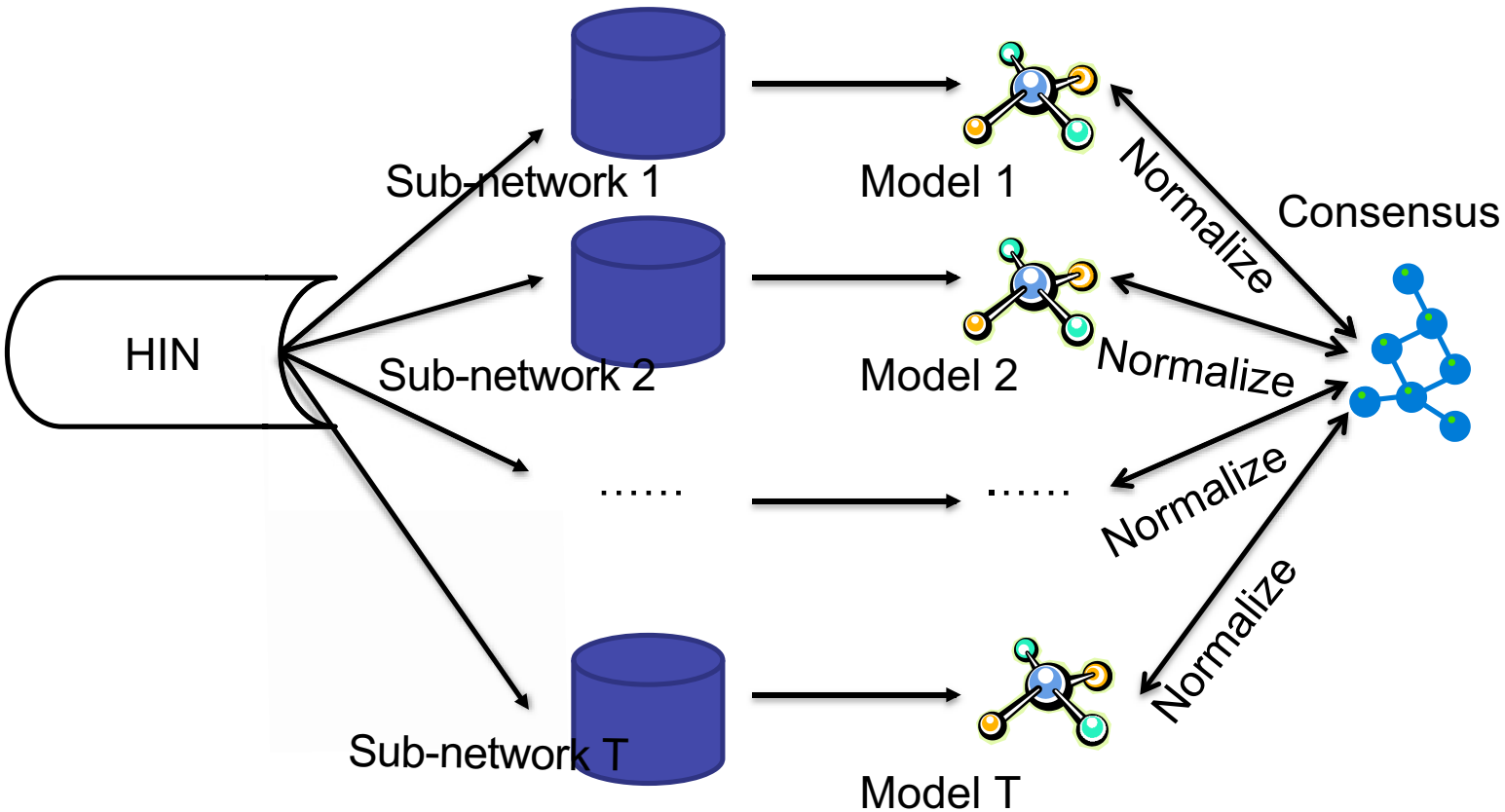
Early studies show that (UQ^{-1}) (or (QV^T)) has the formal properties of conditional probability matrix $[P(w|k)] \in \mathbb{R}_+^{M\times K}$ (or $[P(d,k)]^T \in \mathbb{R}_+^{K\times N}$). This provides theoretical foundation for using NMF to conduct clustering.

Due to this connection, joint NMF has a nice probabilistic interpretation: each element in the matrix V^* is the consensus of $P(d|k)^{(v)}$ weighted by $\lambda_v P(d)^{(v)}$ from different views.

Outline

Extend MultiNMF to HIN

Assume that we are now given T attribute types. Let $\{X^{(1)}, X^{(2)}, \ldots, X^{(T)}\}$ denote the sub-networks, where for each subnework $X^{(t)}$, we have factorizations that $X^{(t)} \approx U^{(t)}(V^{(t)})^T$.



HINMF > MultiNMF + HIN

In HINMF,

1. We expect to get clustering on both center and attribute types at the same time.
2. We wish to learn the strength of each subnetwork automatically.

Objective Function

$$\min_{U^{(t)}s, V^{(t)}s, V^*, \beta^{(t)}s} \sum_{t=1}^{T} \beta^{(t)} \Big(\|X^{(t)} - U^{(t)}(V^{(t)})^T\|_F^2 + \alpha \|V^{(t)}Q^{(t)} - V^*\|_F^2] \Big) \tag{7}$$

$$s.t.\ \forall 1 \leq t \leq T, \quad U^{(t)} \geq 0, V^{(t)} \geq 0, V^* \geq 0,$$

$$\sum_t exp^{-\beta^{(t)}} = 1$$

- We use α as a *fixed* parameter tuning the weight between NMF reconstruction error and the disagreement term.
- $\beta^{(t)}$'s are relative weights of different sub-networks learnt *automatically* from the HIN.

Iterative Update Rules

1. Fixing V^* and $\beta^{(t)}$, minimize over $U^{(v)}$ and $V^{(v)}$:

$$U_{i,k}^{(t)} \leftarrow U_{i,k}^{(t)} \frac{(X^{(t)} V^{(t)})_{i,k} + \alpha \sum_{j=1}^{N} V_{j,k}^{(t)} V_{j,k}^{*}}{(U^{(t)} V^{(t)^T} V^{(t)})_{i,k} + \alpha \sum_{i=1}^{M^{(t)}} U_{i,k}^{(t)} \sum_{j=1}^{N} V_{j,k}^{(t)^2}}$$

$$U^{(t)} \leftarrow U^{(t)} Q^{(t)^{-1}}, \; V^{(t)} \leftarrow V^{(t)} Q^{(t)}$$

$$V_{j,k}^{(t)} \leftarrow V_{j,k}^{(t)} \frac{(X^{(t)^T} U^{(t)})_{j,k} + \alpha V_{j,k}^{*}}{(V^{(t)} U^{(t)^T} U^{(t)})_{j,k} + \alpha V_{j,k}^{(t)}}$$

2. Fixing $U^{(v)}$ and $V^{(v)}$, minimize over V^* and $\beta^{(t)}$:

$$V^* \leftarrow \frac{\sum_{t=1}^{T} \beta^{(t)} V^{(t)} Q^{(t)}}{\sum_{t=1}^{T} \beta^{(t)}} \geq 0, \quad \beta^{(t)} \leftarrow -\log \frac{RE^{(t)}}{\sum_t RE^{(t)}}$$

where $RE^{(t)}$ represents the reconstruction error for the bipartite sub-network related to attribute type t:

$$\|X^{(t)} - U^{(t)}(V^{(t)})^T\|_F^2 + \alpha \|V^{(t)} Q^{(t)} - V^*\|_F^2$$

Obtain Clustering Results

After convergence, the cluster indicators of nodes belonging to the center type can be computed via $\arg\max_k V^*_{j,k}$.

For each attribute type t, cluster nodes of this type indicated by $\arg\max_k U^{(t)}_{i,k} \sum_j V^*_{j,k}$.

This is due to the fact:

$$V^*_{j,k} \approx p(d,k), \quad \sum_j V^*_{j,k} \approx p(k), \quad U^{(t)}_{i,k} \approx p(w|k)$$

Outline

Dataset

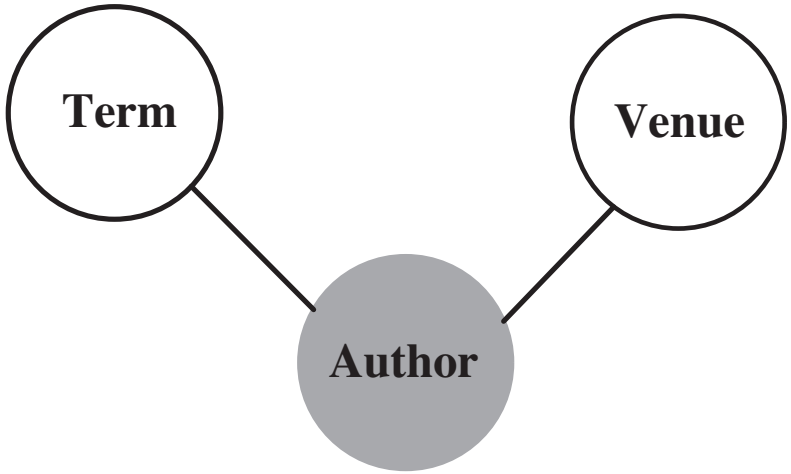


Figure: It is a subset of the DBLP records that belong to four research areas: artificial intelligence, information retrieval, data mining and database. It contains 4023 authors, 20 venues and 11771 unique terms (stop words removed).

Compared Algorithms

We compared with the following algorithms:

- **A-V**: We report the clustering performance after running NMF on the author-venue sub-network.
- **A-T**: It is similar to A-V but we turn to use the author-term sub-network.
- **NetClus**: It is a rank-based algorithm proposed recently by *Sun et al*. to integrate ranking and clustering together in heteroegeneous information networks with star schema.
- **HINMF**: Our proposed method in this paper.

Performance

The accuracy (AC) and normalized mutual information (NMI) are used to measure the performance.

Table: Clustering performance on DBLP dataset (%)

Method	AC(%)		NMI(%)	
	Author	Venue	Author	Venue
A-V	92.35	**100.0**	77.12	**100.0**
A-T	77.24	-	47.28	-
NetClus	90.86	**100.0**	73.51	**100.0**
HINMF	**94.07**	**100.0**	**80.67**	**100.0**

The higher, the better for both *Accuracy* and *Normalized Mutual Information*.

Top Ranked Terms

Besides the evaluation on authors and venues, we list the top ten words for each cluster k by sorting $U^{(2)}_{i,k}$.

Table: Top 10 words in different clusters.

Cluster 1	Cluster 2	Cluster 3	Cluster 4
learning	retrieval	mining	data
based	information	data	database
knowledge	web	clustering	query
problem	search	based	queries
model	query	patterns	xml
algorithm	based	frequent	system
approach	document	large	databases
systems	text	efficient	systems
system	language	databases	based
reasoning	model	classification	processing

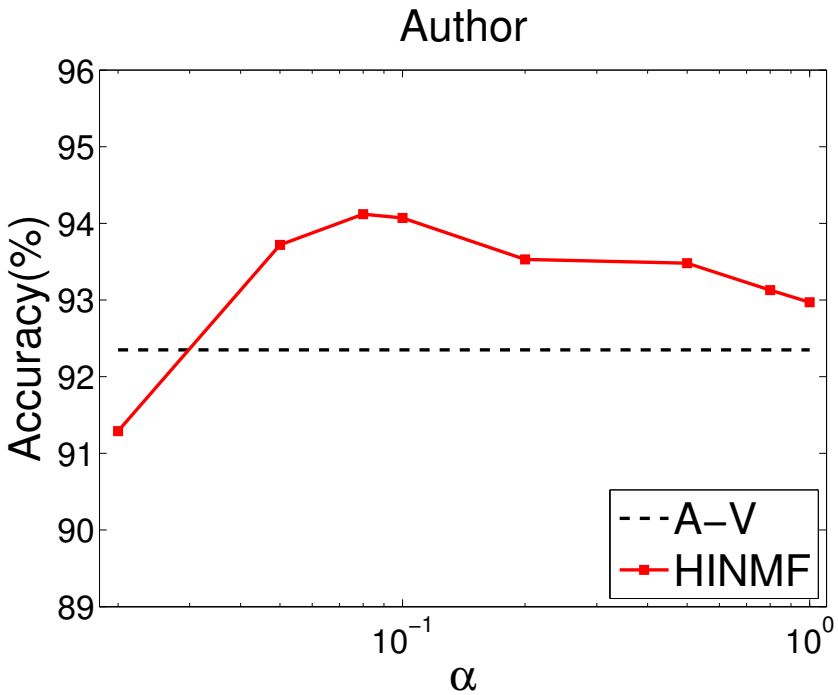
Parameter Study

Recall that

- We use α as a *fixed* parameter tuning the weight between NMF reconstruction error and the disagreement term.
- $\beta^{(t)}$'s are relative weights of different sub-networks learnt *automatically* from the HIN.

Parameter Study

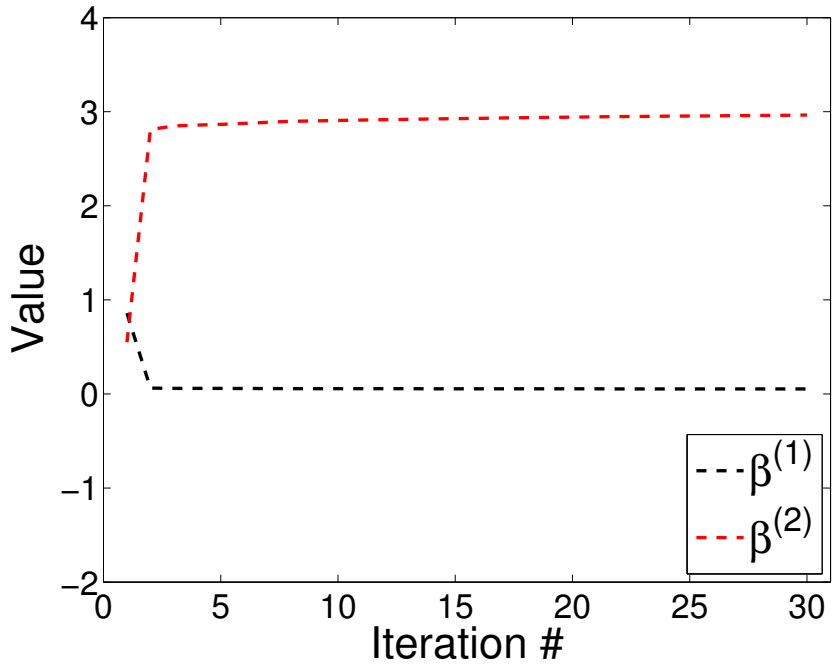
We study the value of α here.



It can be observed that the performance is not much sensitive with respect to different values of α. Thus through the experiment, we set it to be 0.1.

Parameter Study

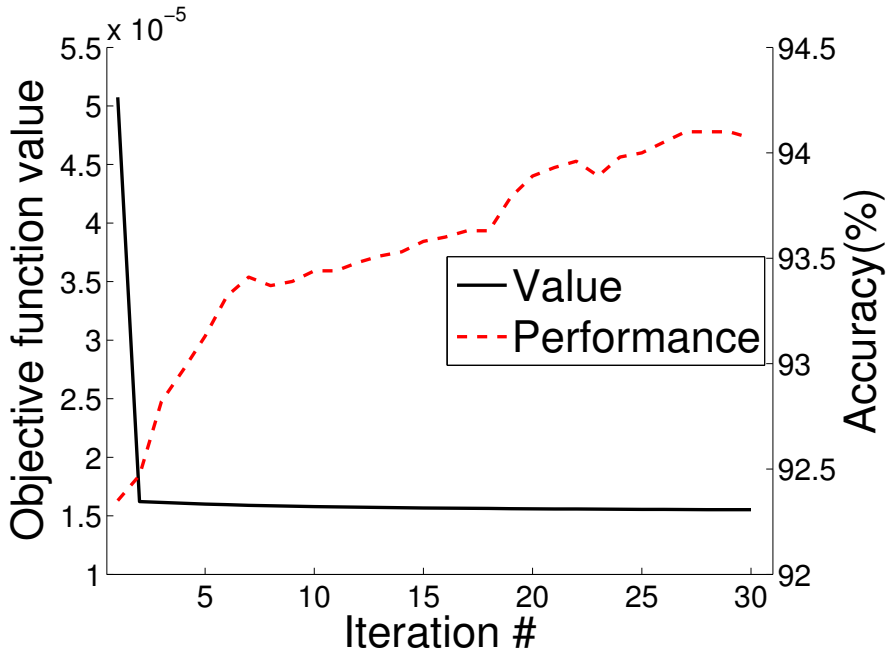
For β, the following figure shows its variation w.r.t. number of iterations.



It is interesting that initially $\beta^{(1)}$ related to author-venue is larger than $\beta^{(2)}$ and the former soon decreases significantly. A possible reason is that during the first several iterations, factorizations learnt on author-term get trapped in the local optimum. By later incorporating the knowledge from author-venue, it gets out of that local minimum.

Convergence Study

It can be proved that the multiplicative update rules are convergent in the paper. Figure below shows the convergence curve together with its performance.



Conclusions

We have proposed an NMF-based approach to solve the HIN clustering problem inspired from Multi-View learning.

- Soft constraints are incorporated.
- Proper normalization is introduced inspired from topic model.
- Automatic strength learning.
- Good clustering result.