Heterogeneous Embedding Propagation for Large-scale E-Commerce User Alignment

Vincent W. Zheng, Mo Sha, Yuchen Li, Hongxia Yang,
Yuan Fang, Zhenjie Zhang, Kian-Lee Tan, Kevin Chen-Chuan Chang

ICDM 2018 @ Singapore

Outline

- **Data and Problem**
- Overall Framework
- Proposed Model
- Experiments
- Conclusion

Data and Problem

Data: User Activity Log

Time	User	IP	Keywords	Auction	Shop
04/05/2017 16:21	PID1	IP2	toys	-	Shop3
04/05/2017 22:12	MID3	IP2	lego	Auction1	Shop2
…	…	…	…	…	…

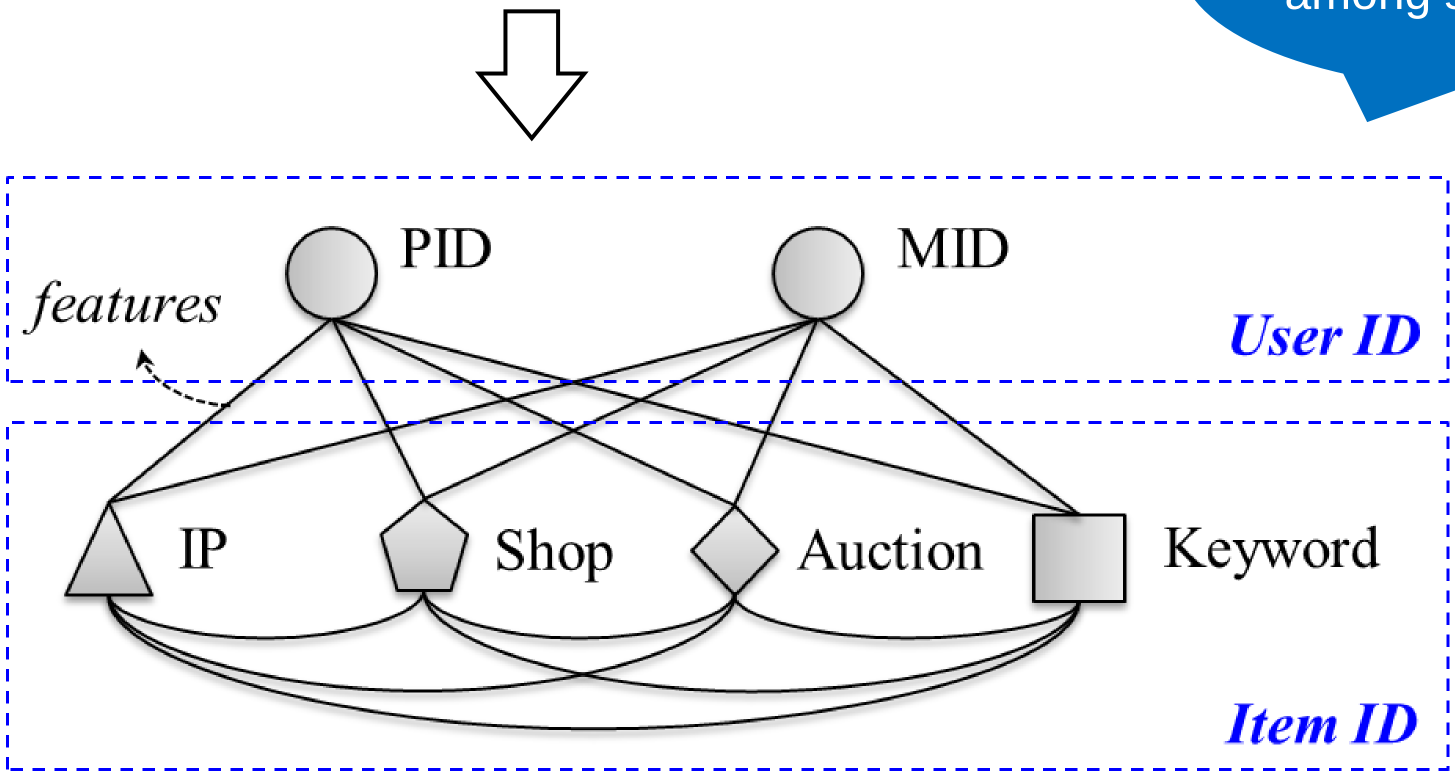
Problem: User ID Linking

To determine if PID1 (a PC identifier) is the same user as MID3 (a mobile device identifier).

Modeling Data as Heterogeneous Interaction Graph

Time	User	IP	Keywords	Auction	Shop
04/05/2017 16:21	PID1	IP2	toys	-	Shop3
04/05/2017 22:12	MID3	IP2	lego	Auction1	Shop2
…	…	…	…	…	…

Exploit interactions among sparse items



Technical challenges: Heterogeneity

Node heterogeneity

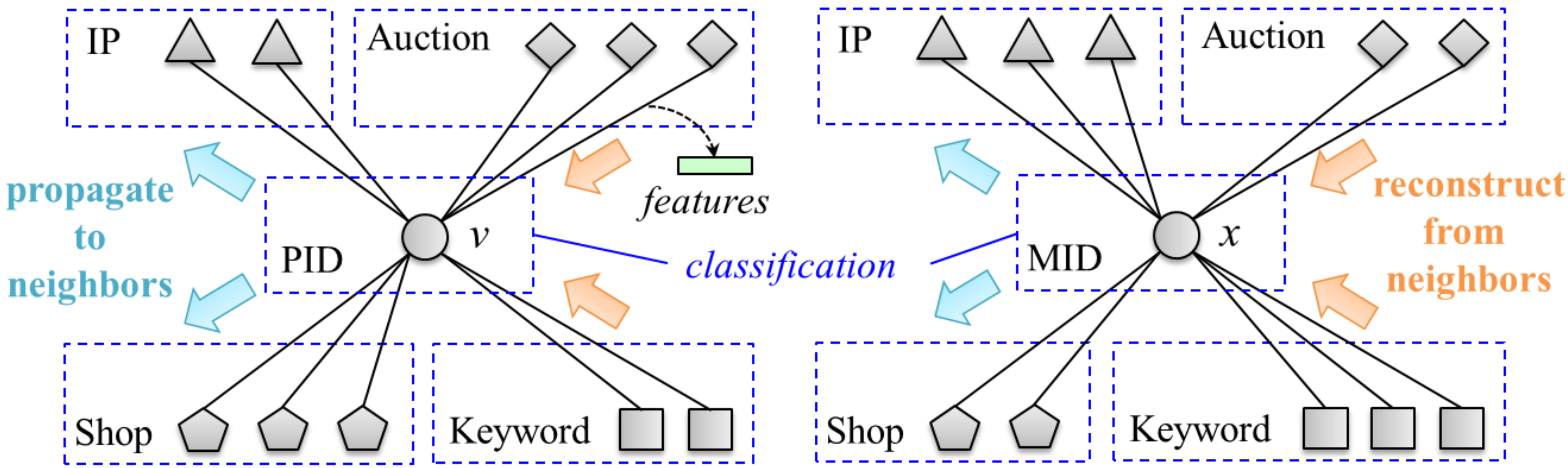
- Different types of nodes with various semantics
  - *Users*
  - *IPs*
  - *Keywords*
  - *Auctions*
  - *Shops*

Edge features

- Time-based historical access patterns
  - *How frequent in past 24 hour?*
  - *How frequent on Sundays?*
  - *How frequent in the evenings?*

Outline

Overall Framework: Heterogeneous Embedding Propagation (HEP)



Classification loss + Reconstruction loss

Outline

Proposed Model: Classification Loss

<u>Semi-supervised learning:</u>

some PID-MID pairs (v_i, u_i) are known to be positive or negative (y_i).

$$P(y_i|v_i, u_i) = \sigma\left(y_i \cdot \mathbf{h}_{v_i}^T W \mathbf{h}_{u_i}\right)$$

$$L_1 = -\frac{1}{m}\sum_{i=1}^{m} \log P(y_i|v_i, u_i)$$

$\mathbf{h}$ is node embedding

W is weight matrix

Proposed Model: Reconstruction Loss

Reconstruction loss

Node embeddings are reconstructed from neighbors, aggregated by node type (c).

$$\tilde{\mathbf{g}}_v^{(c)} = \sum_{u \in N_v^{(c)}} \frac{s_{v,u}}{\sum_{u \in N_v^{(c)}} s_{v,u}} \mathbf{h}_u$$

$$\tilde{\mathbf{g}}_v = \mathrm{CONCAT}\left(\tilde{\mathbf{g}}_v^{(c_1)}, \ldots, \tilde{\mathbf{g}}_v^{(c_{n_1})}\right)$$

$$\tilde{\mathbf{h}}_v = \sigma\left(W'_{\phi(v)} \tilde{\mathbf{g}}_v + \mathbf{b}''_{\phi(v)}\right)$$

$s_{v,u}$ is learnable edge weight (based on edge feature)
$\tilde{\mathbf{h}}$ is reconstructed node embedding
W' and $\mathbf{b}''$ is type specific weight/bias

Proposed Model: Reconstruction Loss

Reconstruction loss

Reconstructed embedding ($\tilde{\mathbf{h}}$) should be close to the target embedding ($\mathbf{h}$).

$$\ell(v,u) = \left[\gamma + \pi(\tilde{\mathbf{h}}_v, \mathbf{h}_v) - \pi(\tilde{\mathbf{h}}_v, \mathbf{h}_u)\right]_+$$

$$L_2 = \frac{1}{|V|} \sum_{v \in V} \sum_{u \sim P_n(u)} \ell(v,u)$$

γ is margin (hyperparameter)
π is distance function between embedding
P_n is negative sampling distribution

Outline

- Data and Problem
- Overall Framework
- Proposed Model
- **Experiments**
- Conclusion

Experiments: Datasets

- Taobao's one-week user activity log in a city
- TB-Top: top 10% active users
- TB-Top: random 10% users

	#record	#PID	#MID	#IP	#shop	#auction	#keyword	#pos	#neg
TB-Top	73,394K	204K	53K	277K	1,125K	3,718K	1,317K	147K	10,611K
TB-Rnd	31,202K	99K	46K	167K	495K	1,082K	437K	57K	2,363K

Experiments: Baselines

- Validating data model (as a graph)
  - FEM: feature engineering
  - LDA: latent Dirichlet allocation
  - GRU: gated recurrent unit
- Validating technical model (HEP)
  - Metapath2vec: meta-path based embedding
  - EP: embedding propagation
  - HEP-: HEP without edge features

Experiments: Results

	TB-Top			TB-Rnd		
	Precision	Recall	F1	Precision	Recall	F1
FEM	60.3	3.4	6.4	68.7	1.9	3.7
LDA	70.4	10.6	18.5	68.3	6.1	11.3
GRU	51.8	26.2	34.8	52.6	22.1	31.2
Metapath2vec	1.7	62.9	3.4	2.3	58.7	4.4
EP	34.3	6.7	11.2	35.0	6.1	10.4
HEP-	32.9	31.3	32.1	34.7	25.0	29.0
HEP	36.5	39.2	**37.8**	44.5	40.5	**42.4**

Non-graph models (FEM, LDA, GRU): high precision but very low recall

HEP: good balance and highest F1

Outline

- Data and Problem
- Overall Framework
- Proposed Model
- Experiments
- **Conclusion**

Conclusion

- Heterogeneous interaction graph
  - Able to capture interactions between items
  - Able to mitigate the sparsity issue
- Heterogeneity challenge
  - Node types
  - Edge features
- Heterogeneous embedding propagation
  - Classification loss
  - Reconstruction loss