北京郵電大學
Beijing University of Posts and Telecommunications

UIC
UNIVERSITY OF ILLINOIS AT CHICAGO

Recent Developments of Deep Heterogeneous Information Network Analysis --Part II Metapath based Data Mining

Chuan Shi

shichuan@bupt.edu.cn

Beijing University of Posts and Telecommunications

Philip S. Yu

psyu@uic.edu

University of Illinois at Chicago

Outline

- ✓ **Metapath based data mining**
  - ✓ Metapath based similarity measure
    - PathSim(VLDB2011), HeteSim(TKDE2014)
  - Metapath based recommendation
    - SemRec(CIKM2015), HeteRec(WSDM2014), SimMF(KAIS2016), FMG(KDD2017)
  - Automatic generation of metapaths
    - RelSim(SDM2016), MP_ESE(TBD2018)
- Heterogeneous information network embedding
- Applications
- Conclusion and future work

Similarity Measure

- Similarity search is important in a wide range of applications.
- The key of similarity search is similarity measure which evaluates the similarity of object pairs.
- Similarity measure is also the base for downstream tasks.

- Web search

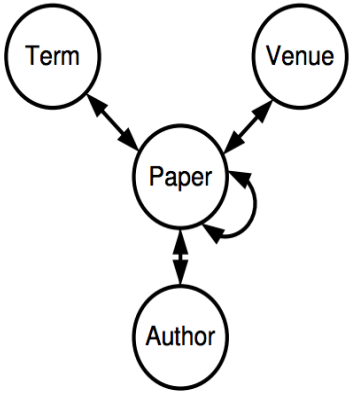
- Product recommendations

How to Define "Similarity" in HIN?

- However, similarity has different semantic meanings under different metapaths in HIN

Table 1: Top-4 similar venues for "DASFAA" under two meta paths.

Rank	$CPAPC$ path	$CPTPC$ path
1	DASFAA	DASFAA
2	DEXA	Data Knowl. Eng.
3	WAIM	ACM Trans. DB Syst.
4	APWeb	Inf. Syst.



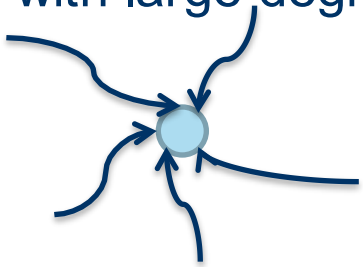
(a) Network Schema

- Limitation of current similarity/proximity measures defined in networks
  - Do NOT distinguish different types of objects and different types of links in the network. E.g., P-PageRank, SimRank

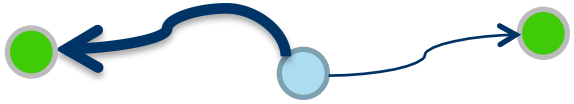
Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, Tianyi Wu. PathSim: Meta Path-Based Top-k Similarity Search in Heterogeneous Information Networks. VLDB, vol.4 pp. 992-1003, 2011

PathSim: Similarity in Terms of "Peers"

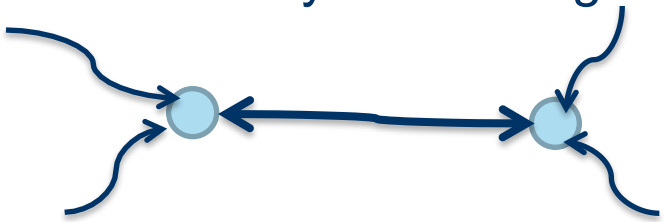
- Path count and Random walk (RW)
  - Favor highly visible objects (objects with large degrees)



- Pairwise random walk (PRW)
  - Favor pure objects (objects with highly skewed scatterness in their in-links or out-links)



- PathSim
  - Favor "peers" (objects with similar visibility and strong connectivity under the given meta path)



Definition of PathSim

- The similarity of two nodes x, y under a symmetric meta path P

$$s(x,y) = \frac{2 \times |\{p_{x \rightsquigarrow y} : p_{x \rightsquigarrow y} \in \mathcal{P}\}|}{|\{p_{x \rightsquigarrow x} : p_{x \rightsquigarrow x} \in \mathcal{P}\}| + |\{p_{y \rightsquigarrow y} : p_{y \rightsquigarrow y} \in \mathcal{P}\}|}$$

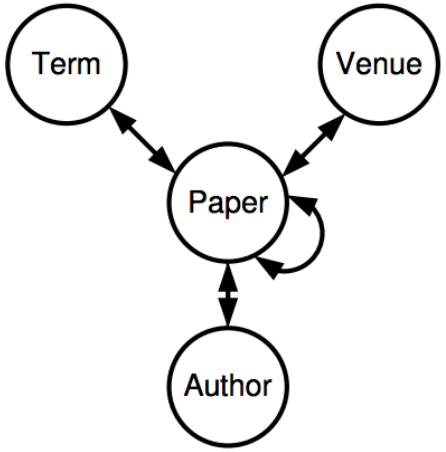
  - Symmetric $s(x_i, x_j) = s(x_j, x_i)$
  - Self-Maximum $s(x_i, x_j) \in [0,1], and\ s(x_i, x_i) = 1$
  - Balance of visibility $s(x_i, x_j) \leq \frac{2}{\sqrt{M_{ii}/M_{jj}} + \sqrt{M_{jj}/M_{ii}}}$

M_{ii} is the number of path instances starting from *i* and ending with *i* following the given meta path

Effectiveness Experiments

Case Study on the query "PKDD" on "DBIS dataset" under meta path CPAPC

Rank	P-PageRank	SimRank	PathSim
1	PKDD	PKDD	PKDD
2	KDD	Local Pattern Detection	ICDM
3	ICDE	KDID	SDM
4	VLDB	KDD	PAKDD
5	SIGMOD	Large-Scale Paral. Data Min.	KDD
6	ICDM	SDM	Data Min. Knowl. Disc.
7	TKDE	ICDM	SIGKDD Expl.
8	PAKDD	SIGKDD Expl.	Knowl. Inf. Syst.
9	SIGIR	Constr.-Bsd. Min. & Induc. DB	J. Intell. Inf. Syst.
10	CIKM	TKDD	KDID



(a) Network Schema

Semantic meanings under different meta paths

le 7: Top-10 most similar authors to "Christos Faloutsos" ler different meta paths on "full DBLP dataset".

(a) Path: APA

Rank	Author
1	Christos Faloutsos
2	Spiros Papadimitriou
3	Jimeng Sun
4	Jia-Yu Pan
5	Agma J. M. Traina
6	Jure Leskovec
7	Caetano Traina Jr.
8	Hanghang Tong
9	Deepayan Chakrabarti
10	Flip Korn

(b) Path: $APCPA$

Rank	Author
1	Christos Faloutsos
2	Jiawei Han
3	Rakesh Agrawal
4	Jian Pei
5	Charu C. Aggarwal
6	H. V. Jagadish
7	Raghu Ramakrishnan
8	Nick Koudas
9	Surajit Chaudhuri
10	Divesh Srivastava

Motivation of HeteSim

- **Similarity search** → **Measure the similarity between same-typed objects**
  - The relatedness measure on different-typed objects is also important.
  - Profile extraction
    - Researcher-related objects
  - Recommender system
    - Users-products
- **Relevance search** → **Measure the relatedness between different-typed objects**

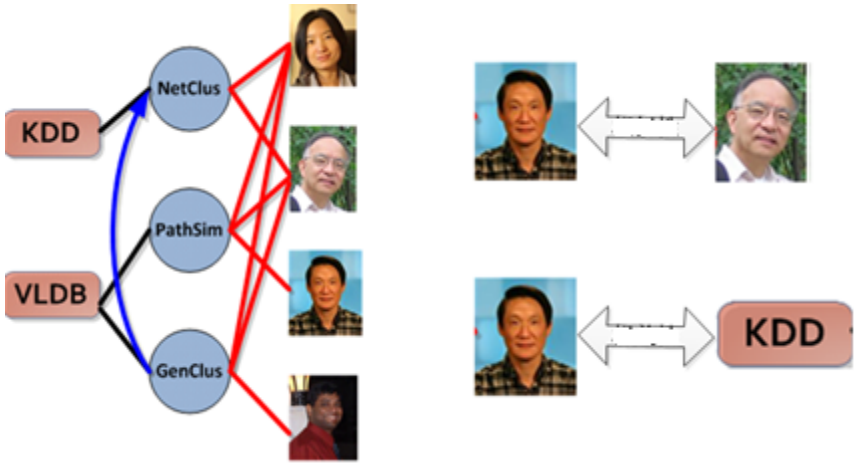
- *It is possible in heterogeneous networks.*

Chuan Shi, Xiangnan Kong, Yue Huang, Philip S. Yu, Bin Wu. HeteSim: A General Framework for Relevance Measure in Heterogeneous Networks. IEEE Transactions on Knowledge and Data Engineering, 2014.

Challenges

- Goal:

  A symmetric measure to evaluate the relevance of arbitrary object pairs



Challenges		Solutions
➢ Path-constrained	⇒	✓ Path-based measure
➢ Symmetric	⇒	✓ Path decomposition

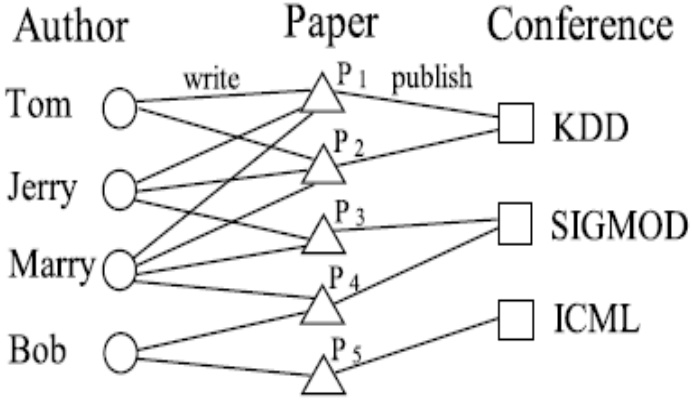
Path-based Measure

- Given a meta path $P = R_1 \circ R_2 \cdots R_l$, HeteSim between *s* and *t* is

$$HeteSim(s,t|R_1 \circ R_2 \circ \cdots \circ R_l) = \frac{1}{|O(s|R_1)||I(t|R_l)|} \sum_{i=1}^{|O(s|R_1)|} \sum_{j=1}^{|I(t|R_l)|} HeteSim(O_i(s|R_1), I_j(t|R_l)|R_2 \circ \cdots \circ R_{l-1})$$

  - Basic idea: Similar objects are related to similar objects
  - Pair-wise random walk in nature

$$HeteSim(Tom, KDD|APC) = \frac{1}{|O(Tom|AP)||I(KDD|PC)|} \sum_{i=1}^{|O(Tom|AP)|} \sum_{j=1}^{|I(KDD|PC)|} HeteSim(O_i(Tom|AP), I_j(KDD|PC))$$



Path Decomposition

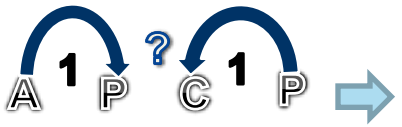
- Two conditions

Even-length Path: meet at middle type

$$HeteSim(s,t|I) = \delta(s,t)$$

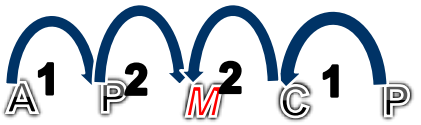
$where\ \delta(s,t) = 1,\ if\ s\ and\ t\ are\ same,\ or\ else\ it\ is\ 0.$

Odd-length Path: meet at atomic relation



$$HeteSim(s,t|R) = HeteSim(s,t|R_O \circ R_I)$$

$$= \frac{1}{|O(s|R_O)||I(t|R_I)|} \sum_{i=1}^{|O(s|R_O)|} \sum_{j=1}^{|I(t|R_I)|} \delta(O_i(s|R_O), I_j(t|R_I))$$



- Decomposition of meta path

$$\boldsymbol{P = (A_1 A_2 \cdot\ \cdot\ \cdot A_{l+1}) = P_L P_R}$$

$$\boldsymbol{P_L = A_1 A_2 \cdot\ \cdot\ \cdot A_{mid-1} M} \qquad \boldsymbol{P_R = M A_{mid+1} \cdot\ \cdot\ \cdot A_{l+1}}$$

Normalization of HeteSim

Before Normalization

$$HeteSim(a, b|\mathcal{P}) = PM_{\mathcal{P}_L}(a,:)PM'_{\mathcal{P}_R^{-1}}(b,:)$$

	A		
A	0.50	0.17	0
	0.17	0.33	0.33
	0	0.33	1

HeteSim(A,A|ABA)

After Normalization

$$HeteSim(a, b|\mathcal{P}) = \frac{PM_{\mathcal{P}_L}(a,:)PM'_{\mathcal{P}_R^{-1}}(b,:)}{\sqrt{\|PM_{\mathcal{P}_L}(a,:)\|\|PM'_{\mathcal{P}_R^{-1}}(b,:)\|}}$$

	A		
A	1	0.41	0
	0.41	1	0.58
	0	0.58	1

HeteSim(A,A|ABA)

- It ranges from 0 to 1.
- The cosine of the probability distributions of a and b reaching the middle type object M.

Comparison of Different Measures

TABLE 1
Comparison of Different Similarity Measures

	Symmetry	Triangle Inequation	Path based	Model	Features
HeteSim	√	×	√	PRW	evaluate relevance of heterogeneous objects based on arbitrary path
PathSim[5]	√	×	√	Path Count	evaluate similarity of same-typed objects based on symmetric path
PCWR[11]	×	×	√	RW	measure proximity to the query nodes based on given path
SimRank[4]	√	×	×	PRW	measure similarity of node pairs based on the similarity of their neighbors
RoleSim[17]	√	√	×	PRW	measure real-valued role similarity based on automorphic equivalence
P-PageRank[3]	×	×	×	RW	measure personalized views of importance based on linkage strcutre

Experiment Results

- Data Sets:
  - *ACM dataset, DBLP dataset, IMDB*
- Case Study on Automatic Profiling and Expert Finding

Automatic Object Profiling Task on Author "Christos Faloutsos" on ACM Data Set

Path	*APVC*		*APT*		*APS*		*APA*	
Rank	Conf.	Score	Terms	Score	Subjects	Score	Authors	Score
1	KDD	0.1198	mining	0.0930	H.2 (database management)	0.1023	Christos Faloutsos	1
2	SIGMOD	0.0284	patterns	0.0926	E.2 (data storage representations)	0.0232	Hanghang Tong	0.4152
3	VLDB	0.0262	scalable	0.0869	G.3 (probability and statistics)	0.0175	Agma Juci M. Traina	0.3250
4	CIKM	0.0083	graphs	0.0816	H.3 (information storage and retrieval)	0.0136	Spiros Papadimitriou	0.2785
5	WWW	0.0060	social	0.0672	H.1 (models and principles)	0.0135	Caetano Traina, Jr.	0.2680

HeteSim		PCRW			
APVC&CVPA		*APVC*		*CVPA*	
Pair	Score	Pair	Score	Pair	Score
C. Faloutsos, KDD	**0.1198**	C. Faloutsos, KDD	0.5517	KDD, C. Faloutsos	**0.0087**
W. B. Croft, SIGIR	**0.1201**	W. B. Croft, SIGIR	0.6481	SIGIR, W. B. Croft	**0.0098**
J. F. Naughton, SIGMOD	**0.1185**	J. F. Naughton, SIGMOD	**0.7647**	SIGMOD, J. F. Naughton	0.0062
A. Gupta, SODA	**0.1225**	A. Gupta, SODA	**0.7647**	SODA, A. Gupta	**0.0090**
Luo Si, SIGIR	0.0734	Luo Si, SIGIR	**0.7059**	SIGIR, Luo Si	0.0030
Yan Chen, SIGCOMM	0.0786	Yan Chen, SIGCOMM	1	SIGCOMM, Yan Chen	0.0013

Applications

- Performances on Query and Clustering

TABLE 5
AUC Values for the Relevance Search of Conferences and Authors Based on Different Paths on DBLP Data Set

Paths	Methods	KDD	ICDM	SDM	SIGMOD	VLDB	ICDE	AAAI	IJCAI	SIGIR
CPA	HeteSim	0.811	0.675	0.950	0.766	0.826	0.732	0.811	0.875	0.613
	PCRW	0.803	0.673	0.939	0.758	0.820	0.726	0.806	0.871	0.606
CPAPA	HeteSim	0.845	0.767	0.715	0.831	0.872	0.791	0.817	0.895	0.952
	PCRW	0.844	0.762	0.710	0.822	0.886	0.789	0.807	0.900	0.949

TABLE 6
Comparison of Clustering Performances for Similarity Measures on DBLP and ACM Data Sets

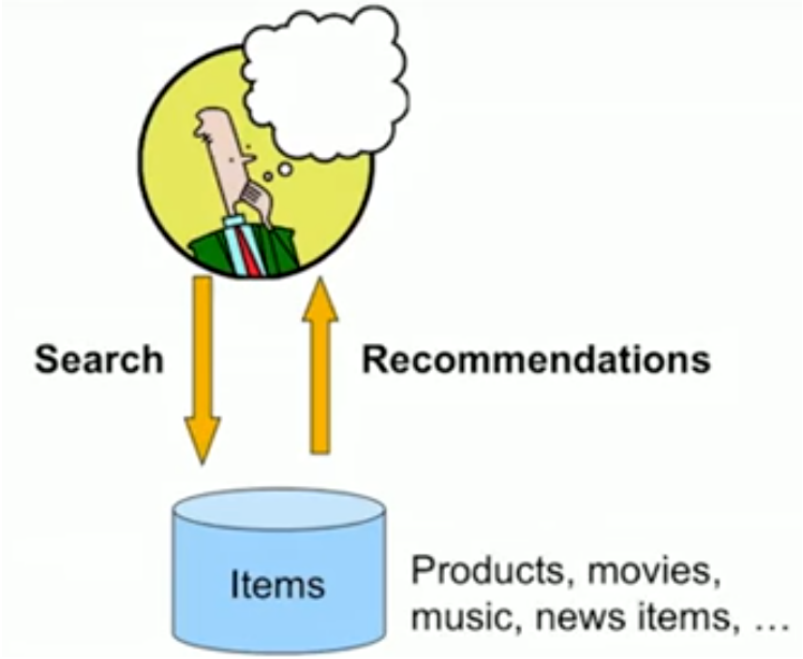
Methods	DBLP dataset						ACM dataset					
	Venue NMI		Author NMI		Paper NMI		Venue NMI		Author NMI		Paper NMI	
	Mean	Dev.	Mean	Dev.	Mean	Dev.	Mean	Dev.	Mean	Dev.	Mean	Dev.
HeteSim	0.768	0.071	**0.728**	0.083	**0.498**	0.067	**0.843**	0.140	0.405	0.1	**0.439**	0.063
PathSim	0.816	0.078	0.672	0.085	0.383	0.058	0.785	0.164	0.378	0.091	0.432	0.087
PCRW	0.709	0.072	0.710	0.080	0.488	0.039	0.840	0.141	**0.414**	0.092	0.429	0.074
SimRank	**0.888**	0.092	0.685	0.066	0.469	0.031	0.835	0.139	0.375	0.115	0.410	0.073
RoleSim	0.278	0.034	0.501	0.040	0.388	0.049	0.389	0.095	0.293	0.016	0.304	0.017
P-PageRank	0.731	0.086	0.441	0.001	0.421	0.063	0.840	0.164	0.363	0.104	0.407	0.093

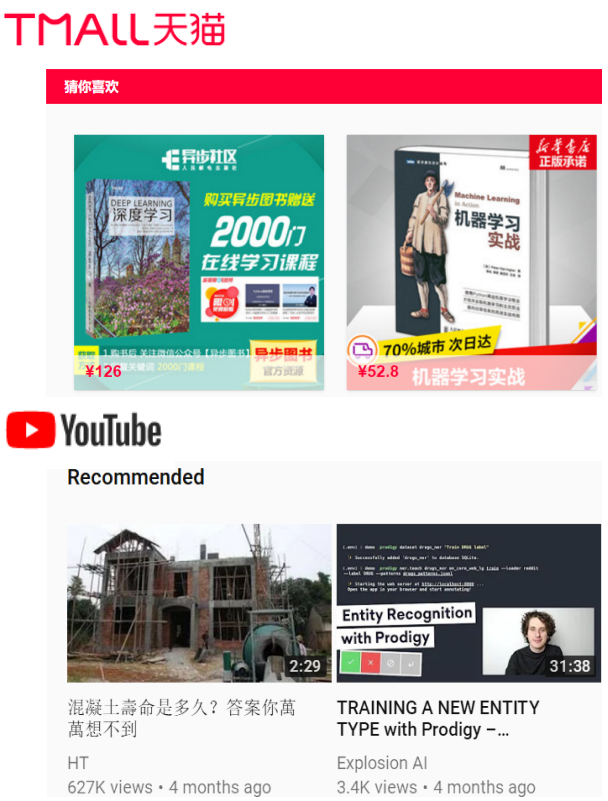
Outline

- ✓ **Metapath based data mining**
  - ✓ Metapath based similarity measure
    - PathSim(VLDB2011), HeteSim(TKDE2014)
  - ✓ Metapath based recommendation
    - SemRec(CIKM2015), HeteRec(WSDM2014), SimMF(KAIS2016), FMG(KDD2017)
  - Automatic generation of metapaths
    - RelSim(SDM2016), MP_ESE(TBD2018)
- Heterogeneous information network embedding
- Applications
- Conclusion and future work

Recommendation System

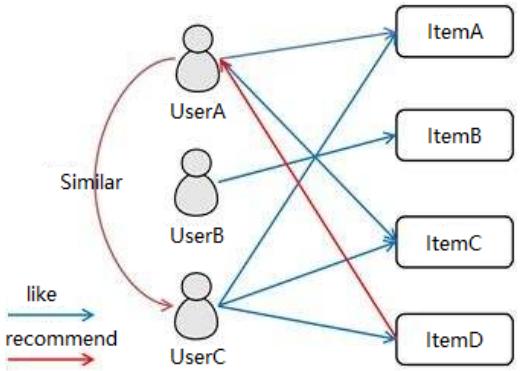
- Recommendation system (RS)
  - Predict the “rating” or “preference” a user would give to an item.
- RS has been widely used in E-commerce and information system.



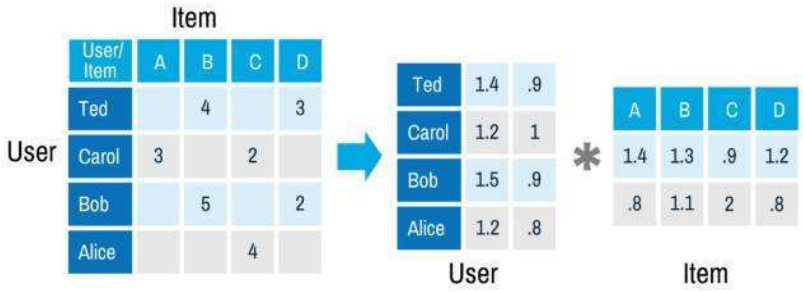


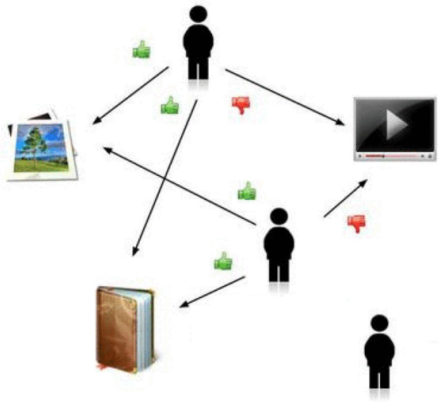
Collaborative Filtering

- Collaborative filtering is a basic recommendation paradigm
  - Similar users have similar behavior
  - Find similar users for recommendation
  - Matrix factorization is a popular CF method



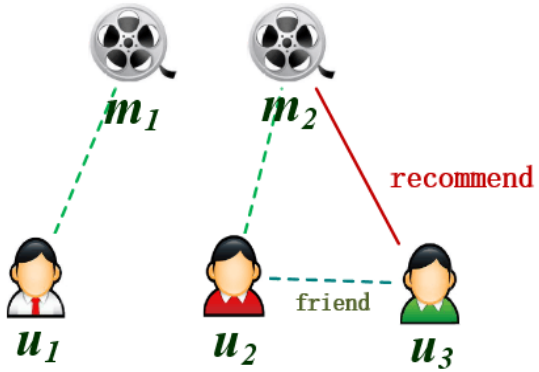
- Disadvantages in collaborative filtering
  - Cold start
  - Sparsity

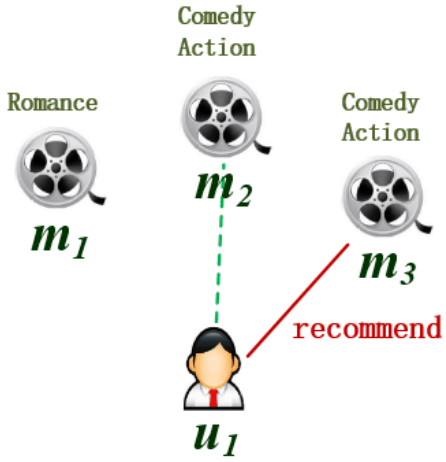




Hybrid Recommendation

- Fusing more information for recommendation
  - Social recommendation: integrate social relations
  - Location based recommendation: integrate location relations
  - Content based recommendation: integrate attribute information

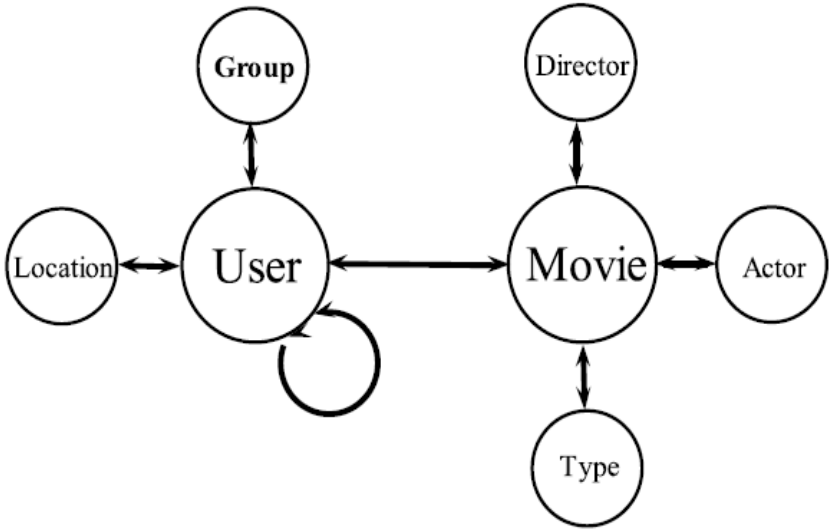




Heterogeneous information network is a more general model to fuse information

Recommendation in HIN

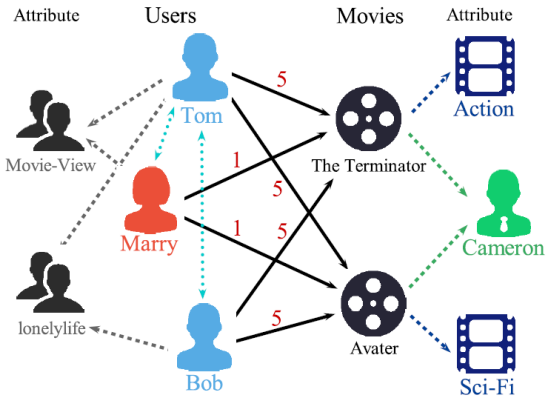
- Organize the objects and relations in recommendation system as a heterogeneous information network.
- Advantages:
  - Integrate information contained in recommendation system
  - Meta path embody rich semantic information

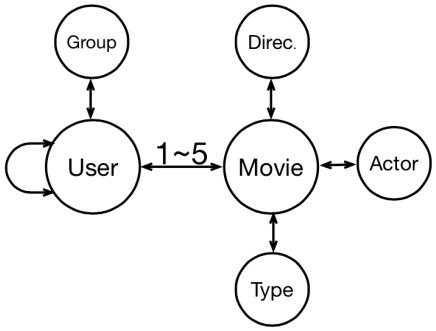


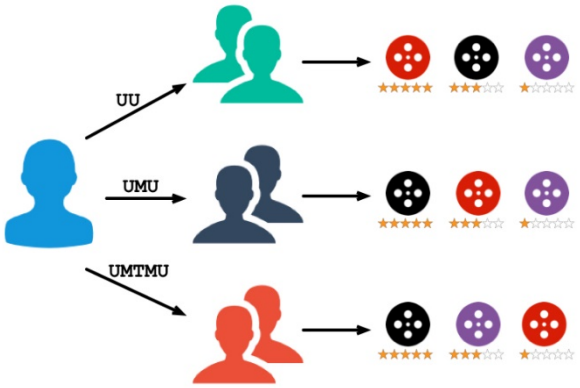
Chuan Shi,Zhiqiang Zhang, Ping Luo, Philip S. Yu, Yading Yue. Semantic Path based Personalized Recommendation on Weighted Heterogeneous Information Networks. CIKM 2015.

Recommendation with Meta Path

- Meta path based recommendation







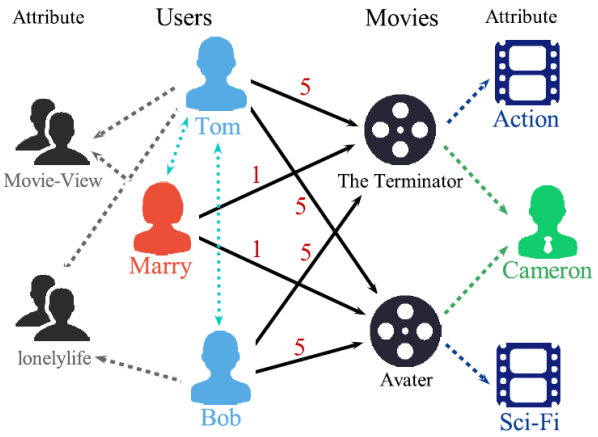
No.	Meta Path	Semantic Meaning	Recommendation Model
1	UU	friends of the target user	Social recommendation
2	UGU	users in the same group of the target user	Member recommendation
3	UMU	users who view the same movies with the target user	Collaborative recommendation
4	UMTMU	users who view movies having the same types with that of the target user	Content recommendation

Challenges: Weight and Combination

Weighted HIN

- Link-constraint Meta Path is a meta path based on a certain constraint on link attribute values.

No.	Link-constrained meta path	Semantic Meaning
1	U(1)M(1)U	Users having the same rating score 1 on some movies as the target user
2	U(i)M(j)U (i=j)	Users having exactly the same rating on some movies as the target user
3	U(i)M(j)U (|i-j|<=1)	users who view movies having the same types with that of the target user



UMU path

$S(Tom, Marry) = S(Bob, Marry)$

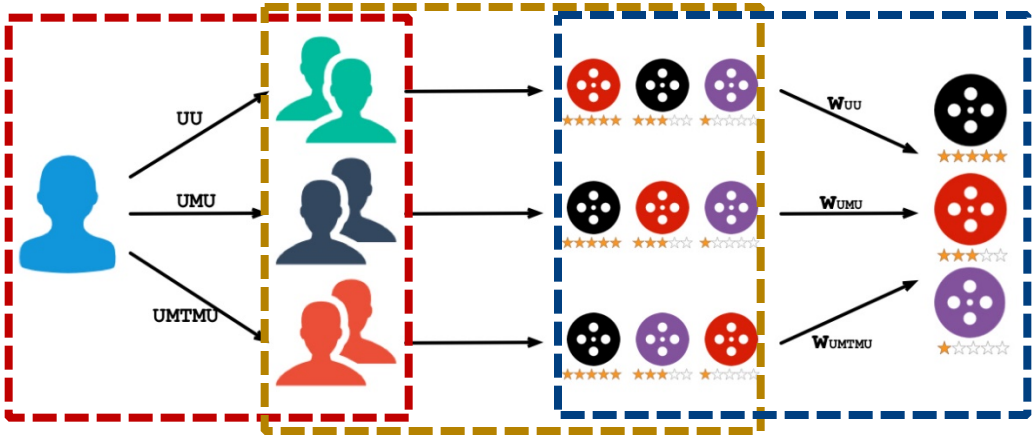
$= S(Tom, Bob)$

U(i)M(j)U (i=j) path

$S(Bob, Tom) > S(Tom, Marry)$

$= S(Bob, Marry)$

SemRec



Similarity Measure

- Find similar users.
- HeteSim [EDBT'12]
- PathSim [VLDB'11]
- PCRW [KDD'10]

Rating Prediction

- Predict rating score along single path

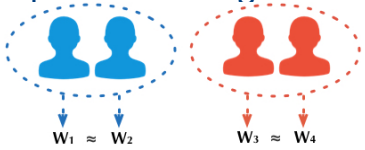
$$Q^{(l)}_{u,i,r} = \sum_v S^{(l)}_{u,v} \times I_{v,i,r}$$

$$I_{v,i,r} = \begin{cases} 1 & R_{v,i} = r \\ 0 & \text{others} \end{cases}$$

$$\hat{R}^{(l)}_{u,i} = \sum_{r=1}^{N} r \times \frac{Q^{(l)}_{u,i,r}}{\sum_{k=1}^{N} Q^{(l)}_{u,i,k}}$$

Weighted Learning

- Combine rating prediction along each path with weights.

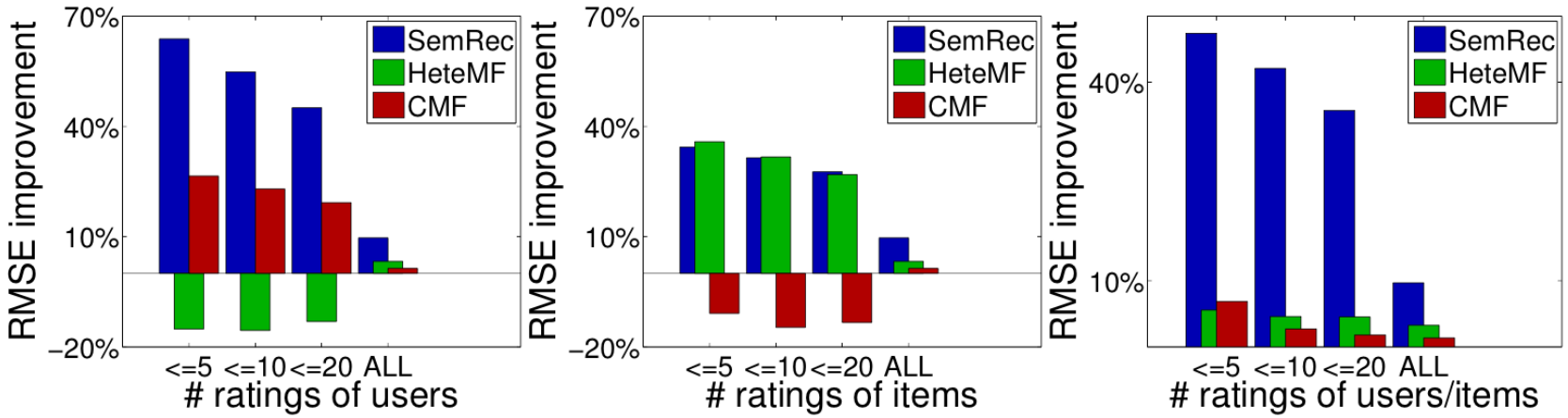


$$\sum_{u=1}^{|U|} \sum_{l=1}^{|\mathcal{P}|} (W^{(l)}_u - \sum_{v=1}^{|U|} \bar{S}^{(l)}_{uv} W^{(l)}_v)^2$$

$$\begin{aligned} \min_W \mathcal{L}_3(W) &= \tfrac{1}{2}||I \odot (R - \sum_{l=1}^{|\mathcal{P}|} diag(W^{(l)})\hat{R}^{(l)})||_2^2 \\ &+ \tfrac{\lambda_1}{2} \sum_{l=1}^{|\mathcal{P}|} ||W^{(l)} - \bar{S}^{(l)} W^{(l)}||_2^2 + \tfrac{\lambda_0}{2}||W||_2^2 \\ s.t. \quad & W \geq 0. \end{aligned}$$

Effectiveness Experiments

Dataset	Training Settings	PMF		SMF		CMF		HeteMF		$\text{SemRec}_{\text{Sgl}}$		$\text{SemRec}_{\text{All}}$		$\text{SemRec}_{\text{Ind}}$		$\text{SemRec}_{\text{Reg}}$	
		RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
Douban	20%	0.9750	0.7198	0.9743	0.7192	0.9285	0.6971	0.8513	0.6342	0.8434	0.6506	0.8125	0.6309	0.8753	0.6412	**0.7844**	**0.6054**
				0.07%	0.08%	4.77%	3.15%	12.69%	11.89%	13.50%	9.61%	16.67%	12.35%	10.23%	10.92%	19.55%	15.89%
	40%	0.8455	0.6319	0.8449	0.6313	0.8273	0.6263	0.7796	0.5927	0.8138	0.6351	0.7814	0.6149	0.8083	0.6032	**0.7452**	**0.5808**
				0.07%	0.09%	2.15%	0.89%	7.79%	6.20%	3.75%	-0.51%	7.58%	2.69%	4.40%	4.54%	11.86%	8.09%
	60%	0.7975	0.6010	0.7967	0.6002	0.8042	0.6090	0.7601	0.5800	0.7937	0.6172	0.7709	0.6098	0.7729	0.5840	**0.7296**	**0.5698**
				0.10%	0.13%	-0.84%	-1.33%	4.69%	3.49%	0.48%	-2.70%	3.34%	-1.46%	3.08%	2.83%	8.51%	5.19%
	80%	0.7673	0.5812	0.7674	0.5815	0.7741	0.5900	0.7550	0.5758	0.7846	0.6142	0.7656	0.6072	0.7540	0.5739	**0.7216**	**0.5639**
				-0.01%	-0.05%	-0.89%	-1.51%	1.60%	0.93%	-2.25%	-5.68%	0.22%	-4.47%	1.73%	1.26%	5.96%	2.98%
Yelp	60%	1.6779	1.2997	1.4843	1.0830	1.6161	1.2628	1.2333	0.9268	1.3252	0.9657	1.2166	0.9040	1.3654	1.0029	**1.2025**	**0.8901**
				11.54%	16.67%	3.68%	2.84%	26.50%	28.69%	21.02%	25.70%	27.49%	30.45%	18.62%	22.84%	28.33%	31.51%
	70%	1.5931	1.2262	1.4017	1.0547	1.5731	1.2224	1.2090	0.9107	1.2889	0.9420	1.1906	0.8873	1.3229	0.9728	**1.1760**	**0.8696**
				12.01%	13.99%	1.26%	0.31%	24.11%	25.73%	19.09%	23.18%	25.27%	27.64%	16.96%	20.67%	26.18%	29.08%
	80%	1.5323	1.1740	1.3678	1.0282	1.5194	1.1740	1.1895	0.8969	1.2576	0.9224	1.1665	0.8723	1.2922	0.9517	**1.1559**	**0.8548**
				10.74%	12.42%	0.84%	0.00%	22.37%	23.60%	17.93%	21.43%	23.87%	25.70%	15.67%	18.94%	24.56%	27.19%
	90%	1.4833	1.1324	1.3377	1.0085	1.4793	1.1405	1.1755	0.8878	1.2331	0.9067	1.1496	0.8616	1.2658	0.9322	**1.1423**	**0.8442**
				9.82%	10.94%	0.27%	-0.72%	20.75%	21.60%	16.87%	19.93%	22.50%	23.91%	14.66%	17.68%	22.99%	25.45%

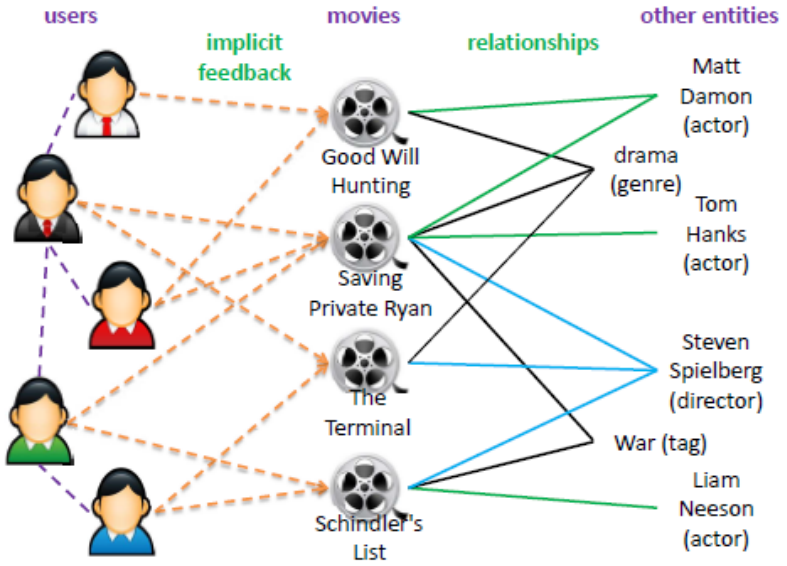


(a) Users (b) Items (c) Users&Items

Motivation of HeteRec

Different types of relationships can be potentially used to improve the recommendation quality.

Diffuse user preferences along **different meta-paths** in information networks to generate latent features for users and items.



Xiao Yu, Xiang Ren, Yizhou Sun, Quanquan Gu. Personalized Entity Recommendation: A Heterogeneous Information Network Approach. WSDM 2014.

Global Recommendation Model

- The user preference diffusion score between user i and item j along metapath $\mathcal{P}$:

$$s(u_i, e_j|\mathcal{P}) = \sum_{e \in \mathcal{I}} \frac{2 \times R_{u_i,e} \times |\{p_{e \rightsquigarrow e_j} : p_{e \rightsquigarrow e_j} \in \mathcal{P}'\}|}{|\{p_{e \rightsquigarrow e} : p_{e \rightsquigarrow e} \in \mathcal{P}'\}| + |\{p_{e_j \rightsquigarrow e_j} : p_{e_j \rightsquigarrow e_j} \in \mathcal{P}'\}|} \quad (2)$$

- Factorize the diffused matrix $\tilde{R}^{(q)}$:

$$(\hat{U}^{(q)}, \hat{V}^{(q)}) = \text{argmin}_{U,V} \|\tilde{R}^{(q)} - UV^T\|_F^2 \quad \text{s.t.} \quad U \geq 0, \; V \geq 0,$$

$\tilde{R}^{(q)}$ is the diffused user preference matrix along the q-th meta-path.

$\hat{U}^{(q)}$ and $\hat{V}^{(q)}$ are the user latent feature and item latent feature respectively.

Personalized Recommendation Model

- Global recommendation scores:

$$r(u_i, e_j) = \sum_{q=1}^{L} \theta_q \cdot \hat{U}_i^{(q)} \hat{V}_j^{(q)T}$$

- Personalized recommendation scores:

$$r^*(u_i, e_j) = \sum_{k=1}^{c} sim(C_k, u_i) \sum_{q=1}^{L} \theta_q^{\{k\}} \cdot \hat{U}_i^{(q)} \hat{V}_j^{(q)T}$$

Apply *k*-means to cluster users into subgroups C

C_k represents user clusters related to target user u_i.

$sim(\cdot,\cdot)$ defines the cosine similarity between the center of cluster C_k and u_i.

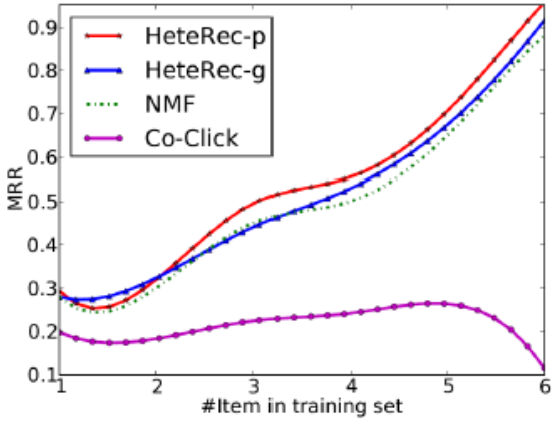
- Bayesian Ranking-Based Optimization:

$$\begin{aligned} O &= -\ln p(\theta|R) = -\ln p(R|\theta)p(\theta) \qquad (9) \\ &= -\sum_{u_i \in \mathcal{U}} \sum_{(e_a > e_b) \in R_i} \ln p(e_a > e_b; u_i|\theta) + \lambda \|\theta\|_2^2 \\ &= -\sum_{u_i \in \mathcal{U}} \sum_{(e_a > e_b) \in R_i} \ln \sigma(r(u_i, e_a) - r(u_i, e_b)) + \lambda \|\theta\|_2^2 \end{aligned}$$

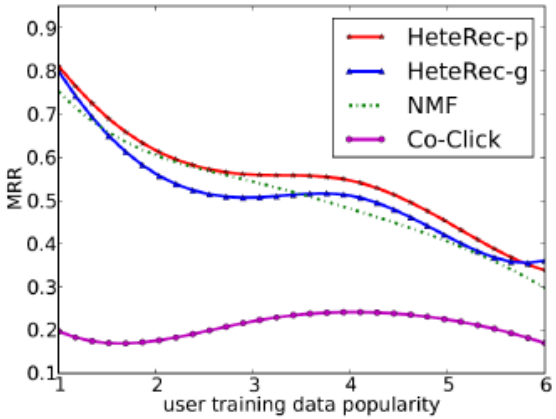
Experiments

Table 3: Performance Comparison

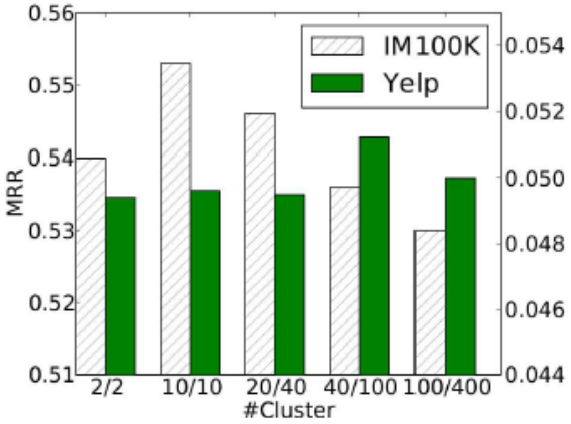
Method	IM100K				Yelp			
	Prec1	Prec5	Prec10	MRR	Prec1	Prec5	Prec10	MRR
Popularity	0.0731	0.0513	0.0489	0.1923	0.00747	0.00825	0.00780	0.0228
Co-Click	0.0668	0.0558	0.0538	0.2041	0.0147	0.0126	0.01132	0.0371
NMF	0.2064	0.1661	0.1491	0.4938	0.0162	0.0131	0.0110	0.0382
Hybrid-SVM	0.2087	0.1441	0.1241	0.4493	0.0122	0.0121	0.0110	0.0337
HeteRec-g	0.2094	0.1791	0.1614	0.5249	0.0165	0.0144	0.0129	0.0422
HeteRec-p	**0.2121**	**0.1932**	**0.1681**	**0.5530**	**0.0213**	**0.0171**	**0.0150**	**0.0513**



(a) Performance Change with User Feedback Number



(b) Performance Change with User Feedback Popularity



(c) Local Model Cluster Number Tuning

Figure 6: Performance Analysis and Parameter Tuning

Motivation of SimMF

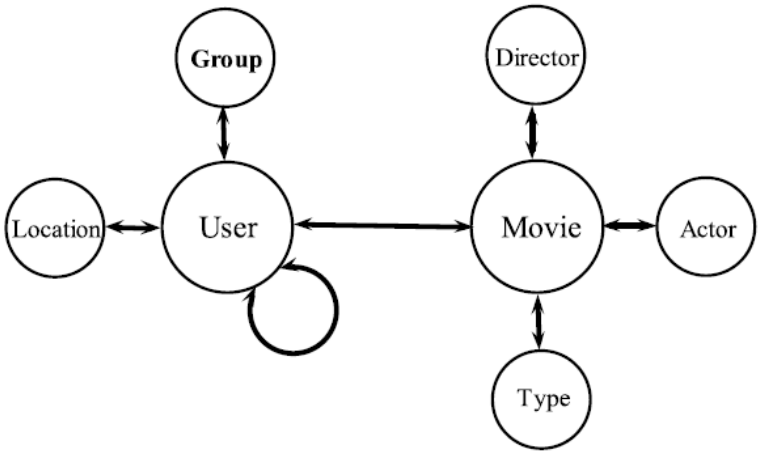
- Social recommendation is an effective method to alleviate data sparsity.
- Limitation of social recommendation
  - Simple similarity information
  - Simple regularization

$$\min_{U,V} \mathcal{J} = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i V_j^T)^2 + \frac{\alpha}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} S_U(i,j)\|U_i - U_j\|^2 + \frac{\lambda_1}{2}(\|U\|^2 + \|V\|^2),$$

Chuan Shi,Jian Liu, Fuzhen Zhuang, Philip S. Yu, Bin Wu. Integrating Heterogeneous Information via Flexible Regularization Framework for Recommendation. KAIS 2016.

Path based Similarity

- Many meta paths can be used to evaluate the similarity of users or items.



User similarity can be evaluated via

"User-User" (UU)
"User-Group-User" (UGU)
"User-Movie-User" (UMU)
.....

Movies similarity can be evaluated via

"Movie-User-Movie"(MUM)
"Movie-Actor-Movie"(MAM)
"Movie-Type-Moive"(MTM)
.....

$$S^{\mathcal{U}} = \sum_l w_l^{\mathcal{U}} S^{(l)} \qquad \Sigma_l w_l^{\mathcal{U}} = 1; 0 \leq w_l^{\mathcal{U}} \leq 1$$
$$S^{\mathcal{I}} = \sum_l w_l^{\mathcal{I}} S^{(l)} \qquad \Sigma_l w_l^{\mathcal{I}} = 1; 0 \leq w_l^{\mathcal{I}} \leq 1$$

Dual Regularization Framework SimMF

- Adopt the similarity of users and items as regularization constraint on the latent factors of users and items.

$$\min_{U,V} \mathcal{L}(R,U,V) = \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{ij}(R_{ij} - U_i V_j^T)^2 + \frac{\alpha}{2} Reg_y^{\mathcal{U}} + \frac{\beta}{2} Reg_y^{\mathcal{I}} + \frac{\lambda_1}{2}\|U\|^2 + \frac{\lambda_2}{2}\|V\|^2$$

$$Reg_{ave}^{\mathcal{U}} = \sum_{i=1}^{m} \|U_i - \frac{\sum_{f\in\mathcal{T}_u^+(i)} S_{if}^{\mathcal{U}} U_f}{\sum_{f\in\mathcal{T}_u^+(i)} S_{if}^{\mathcal{U}}}\|^2 \qquad Reg_{ind}^{\mathcal{U}} = \sum_{i=1}^{m}\sum_{j=1}^{m} S_{ij}^{\mathcal{U}} \|U_i - U_j\|^2$$

$$Reg_{ave}^{\mathcal{I}} = \sum_{j=1}^{n} \|V_j - \frac{\sum_{f\in\mathcal{T}_i^+(j)} S_{jf}^{\mathcal{I}} V_f}{\sum_{f\in\mathcal{T}_i^+(j)} S_{jf}^{\mathcal{I}}}\|^2 \qquad Reg_{ind}^{\mathcal{I}} = \sum_{i=1}^{n}\sum_{j=1}^{n} S_{ij}^{\mathcal{I}} \|V_i - V_j\|^2$$

Experiments

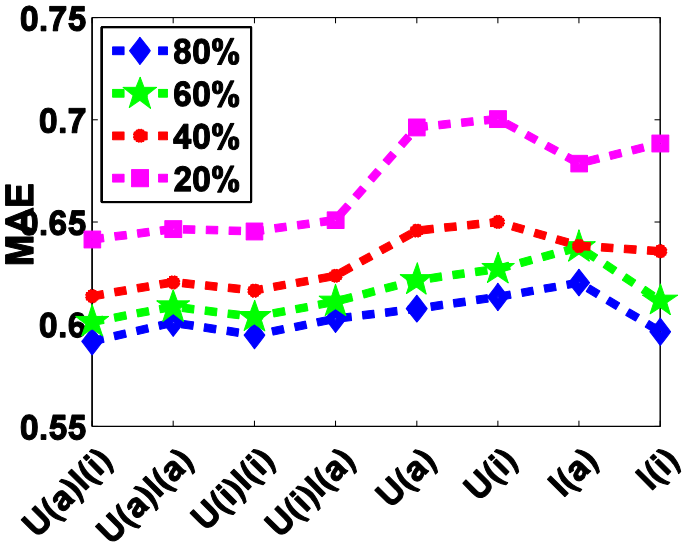
Table 2: Performance Comparisons on Douban (the baseline of improved performances is PMF)

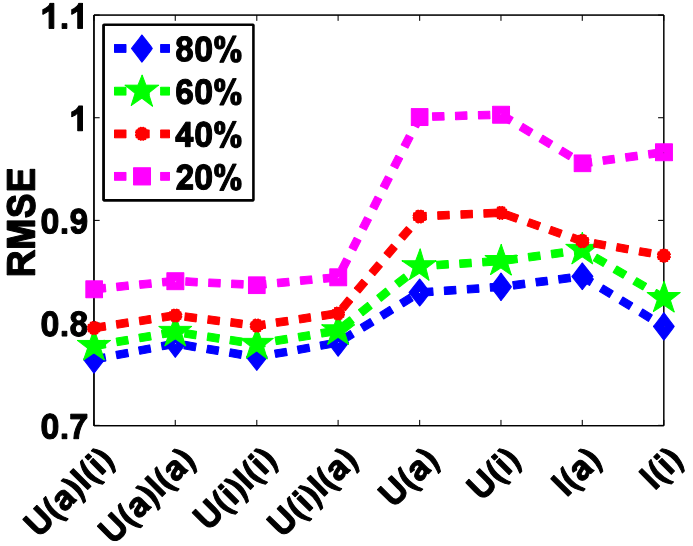
Training	Metrics	UserMean	ItemMean	PMF	Hete-MF	SoMF	SimMF-mean	SimMF-max	SimMF-min
80%	MAE	0.6958	0.6476	0.6325	0.6221	0.6073	0.5974	0.6026	**0.5926**
	Improve	-10.01%	-2.83%		1.64%	3.99%	5.55%	4.73%	6.31%
	RMSE	0.8846	0.8537	0.8815	0.8609	0.8283	0.7729	0.7809	**0.7656**
	Improve	-0.35%	3.15%		2.34%	6.03%	12.32%	11.41%	13.14%
60%	MAE	0.6986	0.6557	0.6591	0.6490	0.6219	0.6060	0.6110	**0.6008**
	Improve	-6.00%	0.35%		1.53%	5.63%	8.06%	7.30%	8.85%
	RMSE	0.8925	0.8748	0.9281	0.9100	0.8584	0.7852	0.7927	**0.7772**
	Improve	3.84%	5.75%		1.95%	7.51%	15.40%	14.59%	16.26%
40%	MAE	0.7052	0.6733	0.7092	0.6933	0.6457	0.6186	0.6237	**0.6134**
	Improve	0.57%	5.07%		2.24%	8.96%	12.77%	12.06%	13.51%
	RMSE	0.9085	0.9139	1.0107	0.9842	0.9034	0.8023	0.8093	**0.7952**
	Improve	10.11%	9.57%		2.62%	10.62%	20.62%	19.93%	21.32%
20%	MAE	0.7227	0.7124	0.8367	0.8235	0.6973	0.6461	0.6509	**0.6417**
	Improve	13.63%	14.85%		1.58%	16.66%	22.78%	22.21%	23.31%
	RMSE	0.9502	1.0006	1.2060	1.1838	1.0037	0.8388	0.8446	**0.8335**
	Improve	21.21%	17.03%		1.84%	16.78%	30.45%	29.97%	30.89%

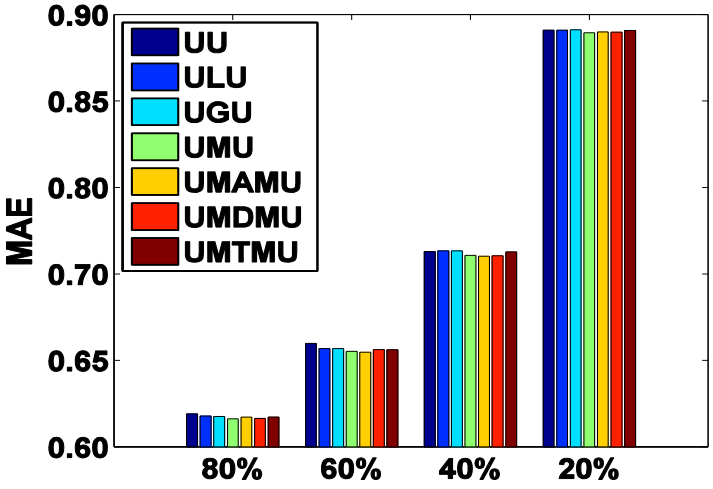
Table 3: Performance Comparisons on Yelp (the baseline of improved performances is PMF)

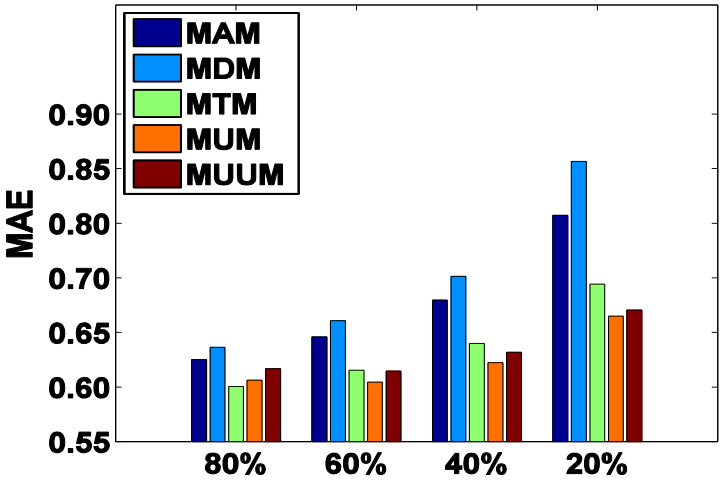
Training	Metrics	UserMean	ItemMean	PMF	Hete-MF	SoMF	SimMF-mean	SimMF-max	SimMF-min
80%	MAE	0.9664	0.8952	1.2201	0.9307	0.8789	0.8292	0.8503	**0.8059**
	Improve	20.79%	26.63%		23.72%	27.96%	32.04%	30.31%	33.95%
	RMSE	1.3443	1.2327	1.6479	1.2773	1.1912	1.0577	1.0708	**1.0465**
	Improve	18.42%	25.20%		22.49%	27.71%	35.82%	35.02%	36.49%
60%	MAE	0.9803	0.9247	1.3835	0.9708	0.9156	0.8366	0.8615	**0.8109**
	Improve	29.14%	33.16%		29.83%	33.82%	39.53%	37.73%	41.39%
	RMSE	1.3556	1.2893	1.8438	1.3352	1.2591	1.0684	1.0842	**1.0532**
	Improve	26.48%	30.07%		27.58%	31.71%	42.05%	41.20%	42.88%
40%	MAE	1.0219	0.9819	1.7081	1.0409	0.9790	0.8509	0.8810	**0.8186**
	Improve	40.17%	42.52%		39.06%	42.68%	50.18%	48.42%	52.18%
	RMSE	1.4241	1.3873	2.2123	1.4343	1.3682	1.0863	1.1031	**1.0686**
	Improve	35.63%	37.29%		35.17%	38.15%	50.90%	50.12%	51.70%
20%	MAE	1.1344	1.1202	2.6935	1.8429	1.1252	0.8687	0.9047	**0.8290**
	Improve	57.88%	58.41%		31.58%	58.23%	67.75%	66.41%	69.22%
	RMSE	1.5958	1.5981	3.2512	2.3357	1.5907	1.1307	1.1733	**1.0944**
	Improve	50.92%	50.85%		28.16%	51.07%	65.22%	63.91%	66.34%

Experiments



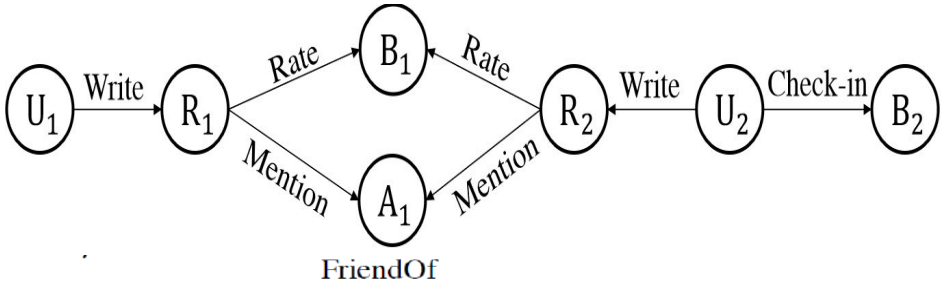






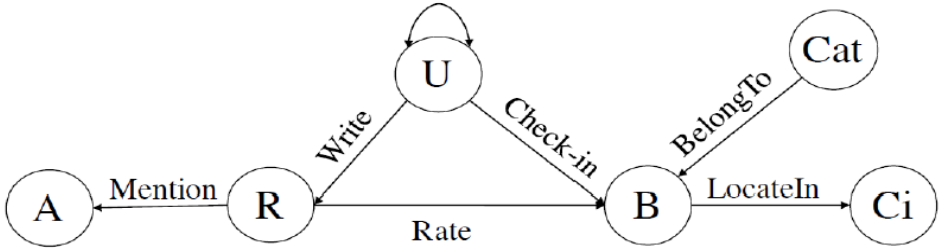
Motivation of FMG

- Recommending strategies can be modeled by meta-paths.
- However, meta-path fails to exploit some complex semantics.



U1 and U2 write the reviews which rate the same bus. **and** mention same asp.

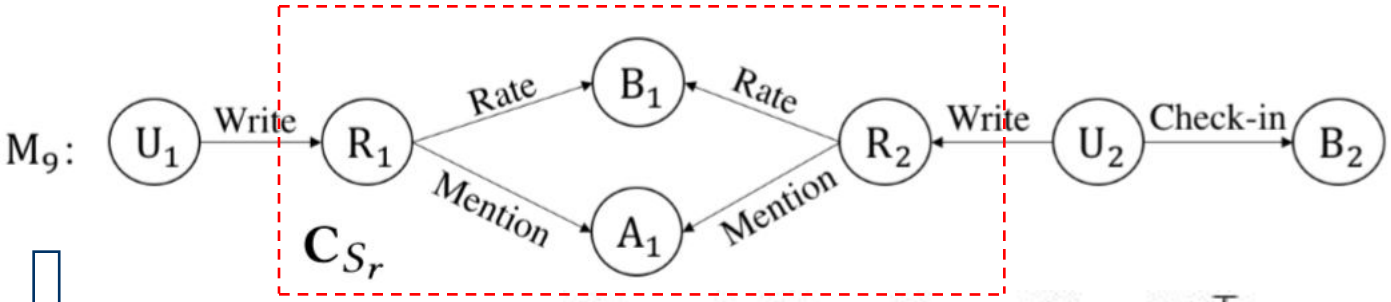
Meta-graph can do it!



Network schema of Yelp

Huan Zhao, Quanming Yao, Jianda Li, Yangqiu Song and Dik Lun Lee. Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks. KDD 2017

Meta-graph based RS



Commuting matrix

$$\text{Compute } \mathbf{C}_{P_1} : \mathbf{C}_{P_1} = \mathbf{W}_{RB} \cdot \mathbf{W}_{RB}^{\top};$$
$$\text{Compute } \mathbf{C}_{P_2} : \mathbf{C}_{P_2} = \mathbf{W}_{RA} \cdot \mathbf{W}_{RA}^{\top};$$
$$\text{Compute } \mathbf{C}_{S_r} : \mathbf{C}_{S_r} = \mathbf{C}_{P_1} \odot \mathbf{C}_{P_2};$$
$$\text{Compute } \mathbf{C}_{M_9} : \mathbf{C}_{M_9} = \mathbf{W}_{UR} \cdot \mathbf{C}_{S_r} \cdot \mathbf{W}_{UR}^{\top} \cdot \mathbf{W}_{UB}.$$

Assemble **L** meta-graphs

“MF+FM” framework

Matrix Factorization

Factorization Machine
+group lasso(FMG)

Optimization Framework

- For each meta-graph, do MF:

$$\min_{\mathbf{U},\mathbf{B}} \frac{1}{2}||P_\Omega(\mathbf{U}\mathbf{B}^\top - \mathbf{R})||_F^2 + \frac{\lambda_u}{2}||\mathbf{U}||_F^2 + \frac{\lambda_b}{2}||\mathbf{B}||_F^2,$$

$$\mathbf{x}^n = \underbrace{\mathbf{u}_i^{(1)}, \ldots, \mathbf{u}_i^{(l)}, \ldots, \mathbf{u}_i^{(L)}}_{L\times F} \underbrace{\mathbf{b}_j^{(1)}, \ldots, \mathbf{b}_j^{(l)}, \ldots, \mathbf{b}_j^{(L)}}_{L\times F},$$

- Do FM(rating for x^n):

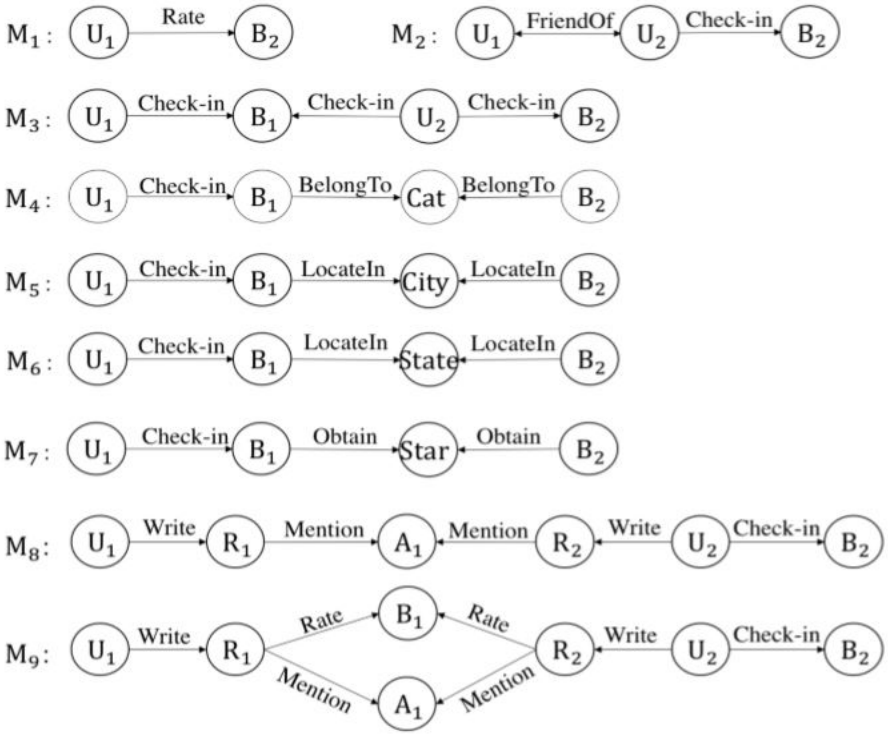
$$\hat{y}^n(\mathbf{w},\mathbf{V}) = w_0 + \sum_{i=1}^{d} w_i x_i^n + \sum_{i=1}^{d}\sum_{j=i+1}^{d} \langle \mathbf{v}_i, \mathbf{v}_j \rangle x_i^n x_j^n,$$

- Objective function:

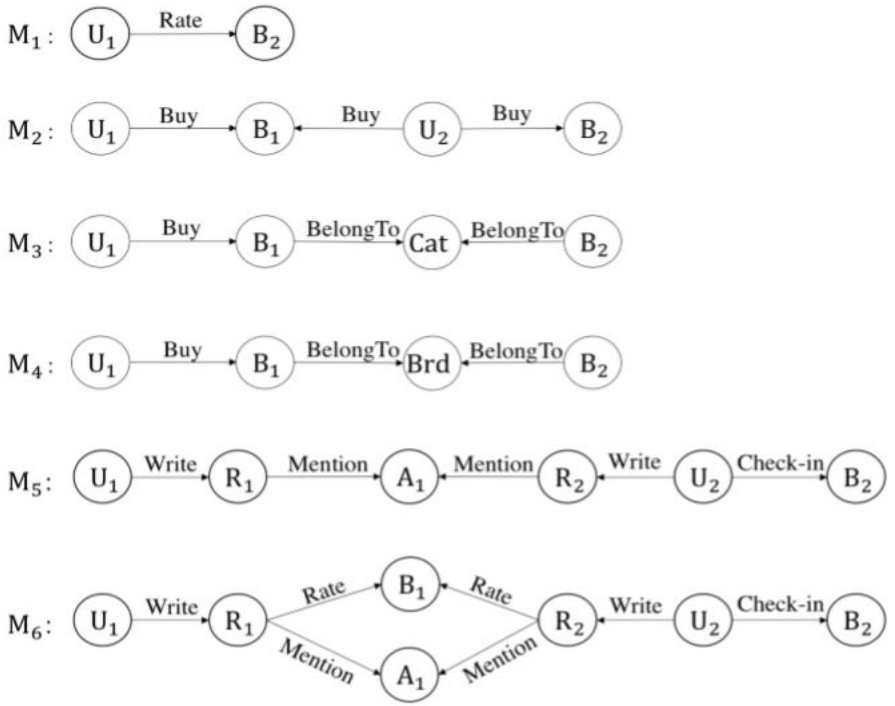
$$h(\mathbf{w},\mathbf{V}) = \sum_{n=1}^{N} (y^n - \hat{y}^n(\mathbf{w},\mathbf{V}))^2 + \lambda_w \Phi_{\mathbf{w}}(\mathbf{w}) + \lambda_v \Phi_{\mathbf{V}}(\mathbf{V})$$

$$\Phi_{\mathbf{w}}(\mathbf{w}) = \sum_{l=1}^{2L} ||\mathbf{w}_l||_2, \quad \Phi_{\mathbf{V}}(\mathbf{V}) = \sum_{l=1}^{2L} ||\mathbf{V}_l||_F,$$

Experiments



(a) Yelp-200K (Star: the average stars a business obtained).



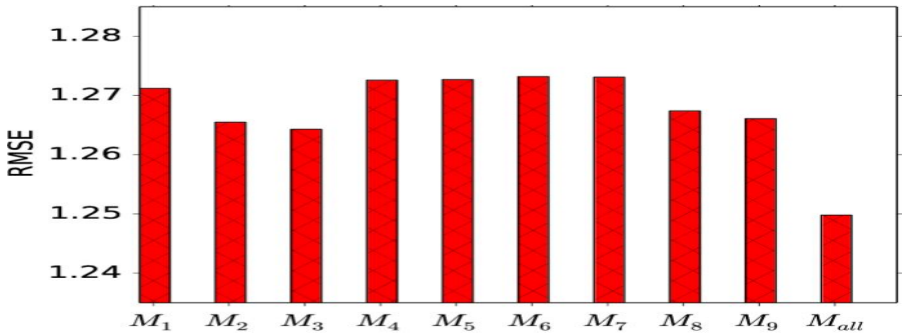
(b) Amazon-200K (Brd: brand of the item).

Figure 3: Meta-graphs used for Amazon and Yelp datasets.

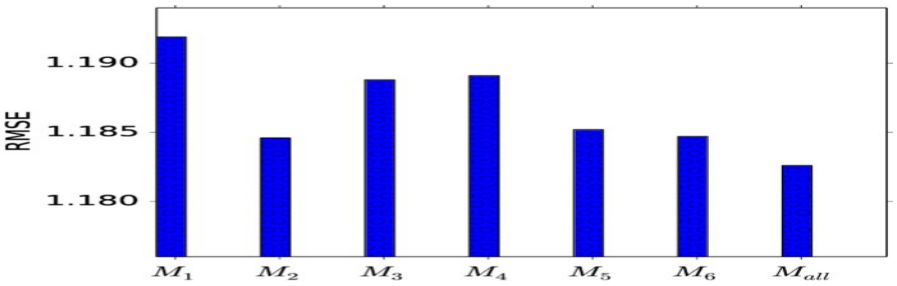
Experiments

Table 3: Recommending performance in terms of RMSE. Percentages in the brackets are the reduction of RMSE comparing our approach with the corresponding approaches in the table header.

	Amazon-200K	Yelp-200K	CIKM-Yelp	CIKM-Douban
RegSVD	2.9656 (+60.0%)	2.5141 (+50.5%)	1.5323 (+27.7%)	0.7673 (+9.0%)
FMR	1.3462 (+11.9%)	1.7637 (+29.4%)	1.4342 (+22.8%)	0.7524 (+7.2%)
HeteRec	2.5368 (+53.2%)	2.3475 (+47.0%)	1.4891 (+25.6%)	0.7671 (+9.0%)
SemRec	- -	1.4603 (+14.7%)	1.1559 (+4.2%)	0.7216 (+3.2%)
FMG	**1.1864**	**1.2456**	**1.1074**	**0.6985**



(a) Yelp.



(b) Amazon.

Figure 5: RMSE of single meta-graph on Yelp and Amazon datasets. M_{all} is our model trained with all meta-graphs.

Outline

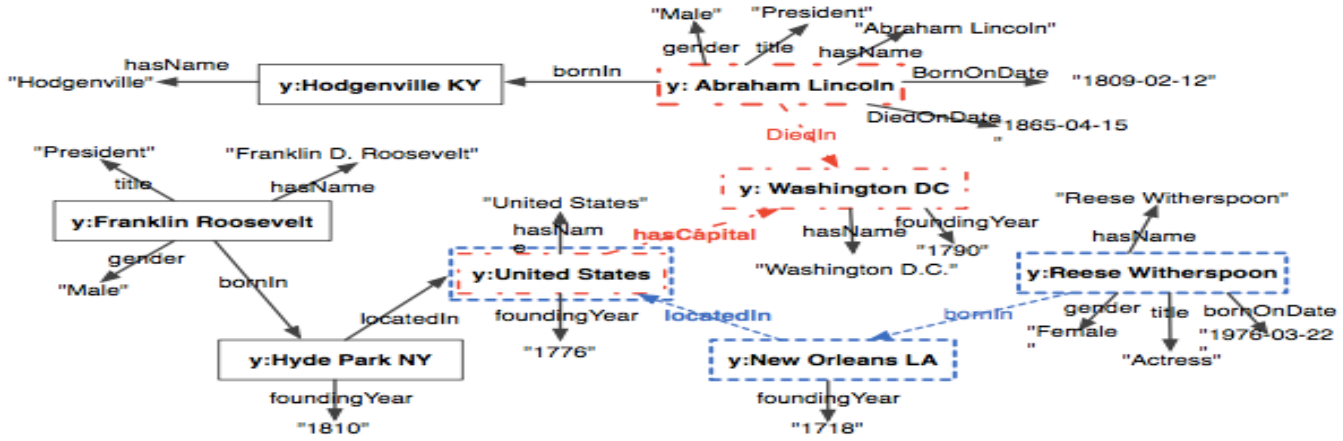
- ✓ **Metapath based data mining**
  - ✓ Metapath based similarity measure
    - Pathsim(VLDB2011), HeteSim(TKDE2014)
  - ✓ Metapath based recommendation
    - SemRec(CIKM2015), HeteRec(WSDM2014), SimMF(KAIS2016), FMG(KDD2017)
  - ✓ **Automatic generation of metapaths**
    - RelSim(SDM2016), MP_ESE(TBD2018)
- Heterogeneous information network embedding
- Applications
- Conclusion and future work

Schema-rich HIN

- Previous work assume that metapaths can be easily given.
- It is not true for some complex HIN, e.g. KG
- Knowledge Graph (KG) contains rich knowledge facts.

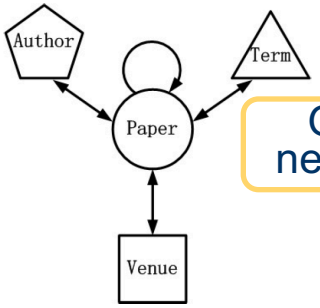
<Subject, Property, Object> ➡ *<Node, Link, Node>*

- Knowledge graph can be seen as a schema-rich HIN
  - Types of nodes and relations are huge
  - Network schema are complex



Challenges

- Simple HIN

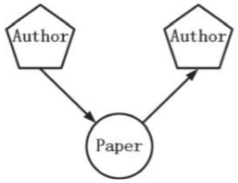


Given simple network schema

Task

Find similar authors working together on the same paper

Provides meta paths



Author-Paper-Author

Search

DBLP network

- Schema-rich HIN

Given **complex** network schema

Task

Find similar person serving the same party

Cannot provides meta paths

?

?

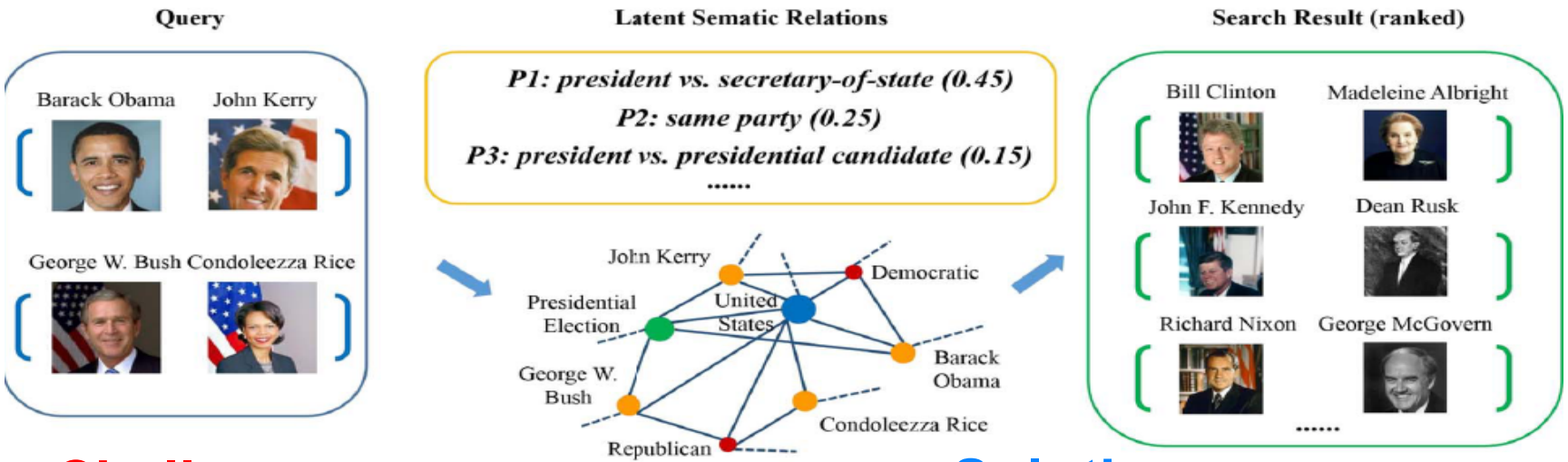
Search

yago
select knowledge

Knowledge graph Yago

Motivation of RelSim

- Relation Similarity Search
  - User provides a set of simple examples
  - Automatically detect the the latent semantic relation (LSR)



Challenges

- The similarity between relation instances
- Latent semantic relation detection

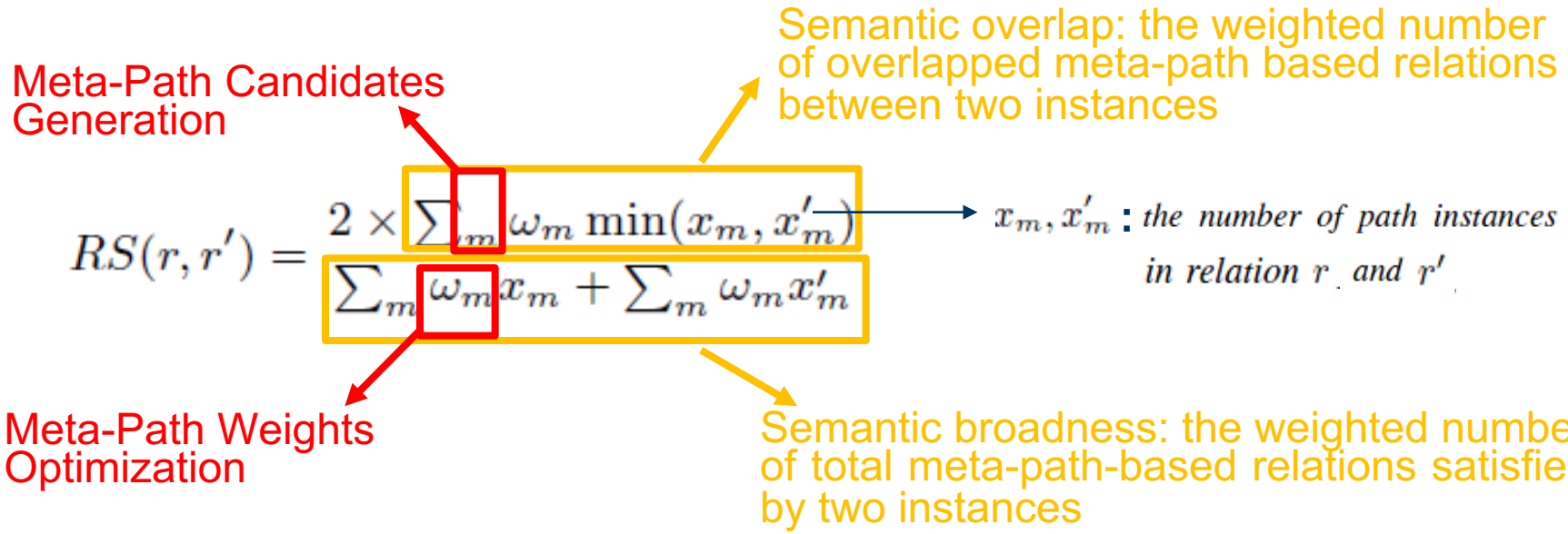
Solutions

- ✓ A meta-path-based relation similarity measure
- ✓ Meta path generation
- ✓ Meta path weights optimization

Chenguang Wang, Yizhou Sun, Yanglei Song, Jiawei Han, Yangqiu Song, Lidan Wang, Ming Zhang. RelSim: Relation Similarity Search in Schema-Rich Heterogeneous Information Networks. SDM 2016.

RelSim

- Given an LSR $\{w_m, \mathcal{P}_m\}_{m=1}^{M}$, RelSim between two relation instances (r and r') is defined as

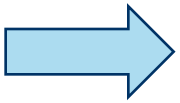


- Intuition: two relation instances are more similar when sharing more important (heavily weighted) meta-paths
- Properties: Range, Symmetric, Self-maximum

- Query-based meta-path generation algorithm (QMPG)
  - Given user query Q, generate meta paths connecting relation instances.
  - Combine these meta-paths to construct a simple schema

1,500+ entity types

35,000+ relation types



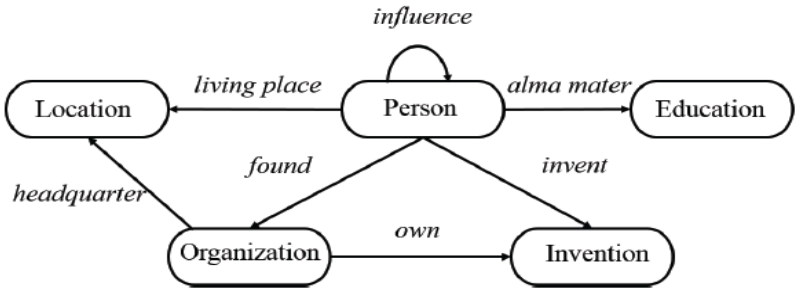


Figure 2: The query-based network schema for query $Q = \{\langle$Larry Page, Sergey Brin$\rangle, \langle$Jerry Yang, David Filo$\rangle\}$.

- Optimization model:

Important" meta-paths have higher weights, while "unimportant" ones near 0

Hinge loss function

$$\min_{\omega} \sum_{k=1}^{K} \max\{0, c - \omega^T \mathbf{x}_k + \omega^T \tilde{\mathbf{x}}_k\}$$

$$\text{s.t.} \quad \omega_m \geq 0 \quad \forall m = 1, \cdots, M$$

$$\sum_{m=1}^{M} \omega_m = 1$$

Experiment Results

- Data Sets:
  - Five popular relation categories in Freebase

Table 1: **Rel-Full** dataset statistics. #Entities means the number of entities; #Relations means the number of relations.

Relation Categories	#Entities	#Relations	Examples
⟨*Organization, Founder*⟩	9,836,649	560,688,893	⟨Google, Larry Page⟩, ⟨Microsoft, Bill Gates⟩, ⟨Facebook, Mark Zuckerberg⟩
⟨*Book, Author*⟩	16,640,478	981,788,232	⟨Gone with the Wind, Margaret Mitchell⟩, ⟨The Kite Runner, Khaled Hosseini⟩
⟨*Actor, Film*⟩	4,340,986	182,121,412	⟨Leonardo DiCaprio, Inception⟩, ⟨Daniel Radcliffe, Harry Potter⟩, ⟨Jack Nicholson, Head⟩
⟨*Location, Contains*⟩	1,037,791	62,229,669	⟨United States of America, New York⟩, ⟨Victoria, Chillingollah⟩, ⟨New Mexico, Davis House⟩
⟨*Music, Track*⟩	1,653,931	86,658,343	⟨My Worlds, Baby⟩, ⟨21, Someone Like You⟩, ⟨Thriller, Beat It⟩
Total	26,841,657	1,483,834,223	⟨Google, Larry Page⟩, ⟨Leonardo DiCaprio, Inception⟩, ⟨Thriller, Beat It⟩

- Effectiveness:

Table 2: Performance (NDCG@K) of relation similarity search on **Rel-Full**.

	NDCG@5	NDCG@10	NDCG@20
VSM-S	0.5389	0.6296	0.7225
LRA-S	0.5880	0.6848	0.7814
IW-S	0.5210	0.6095	0.7010
RelSim-S	0.6395	0.7427	0.8432
RelSim-WS	**0.6651**	**0.7716**	**0.9559**

Experiment Results

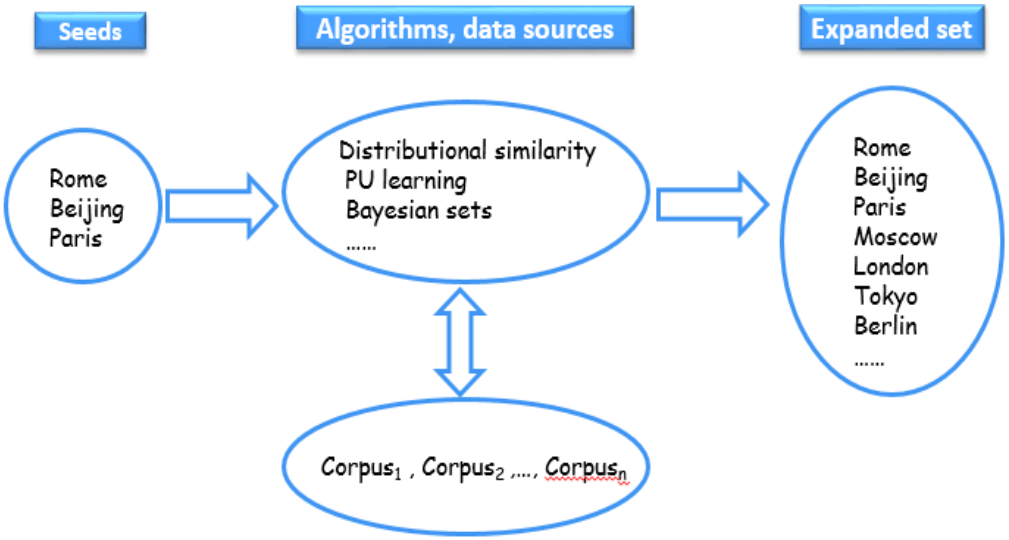
- Case study of query-based meta-paths

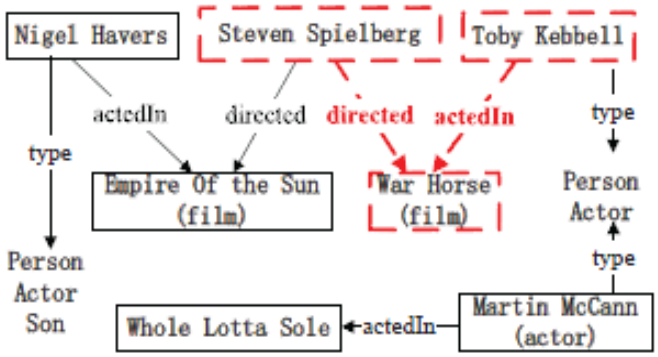
Table 4: Example query-based meta-paths on **Rel-Full**. We show the most important four query-based meta-paths of different queries

Query: {⟨Google, Larry Page⟩, ⟨Microsoft, Bill Gates⟩, etc.}	ω
$Organization \xrightarrow{\text{is founded by}} Founder$	0.384
$Organization \xrightarrow{\text{run business in}} Industry \xrightarrow{\text{win award in}^{-1}} Founder$	0.274
$Organization \xrightarrow{\text{is founded by}} Person \xrightarrow{\text{is influence peer}^{-1}} Founder$	0.174
$Organization \xrightarrow{'\text{s leadership}} Person \xrightarrow{\text{mailing address}} Location \xrightarrow{\text{mailing address}^{-1}} Founder$	0.115
Query: {⟨Google, Larry Page⟩, ⟨Yahoo!, Marissa Mayer⟩, etc.}	ω
$Organization \xrightarrow{\text{run by}} CEO \xrightarrow{\text{job title}^{-1}} Founder$	0.320
$Organization \xrightarrow{\text{founded date}} Date \xrightarrow{\text{graduation date}^{-1}} Founder$	0.229
$Organization \xrightarrow{\text{headquarter}} Location \xrightarrow{\text{education institute}^{-1}} Founder$	0.207
$Organization \xrightarrow{\text{run business in}} Industry \xrightarrow{\text{win award in}^{-1}} Founder$	0.113

Entity Set Expansion in HIN

- Entity set expansion (ESE)
  - Expanding a small set into a more complete set





Can we do ESE in knowledge graph?

Yuyan Zheng, Chuan Shi, Xiaohuan Cao, Xiaoli Li, Bin Wu. A Meta Path based Method for Entity Set Expansion in Knowledge Graph. TBD 2018.

Basic idea of MP_ESE

- Basic idea
  - Seeds:

{Toby Kebbell, Nigel Havers, Harrison Ford}

  - Important meta path:

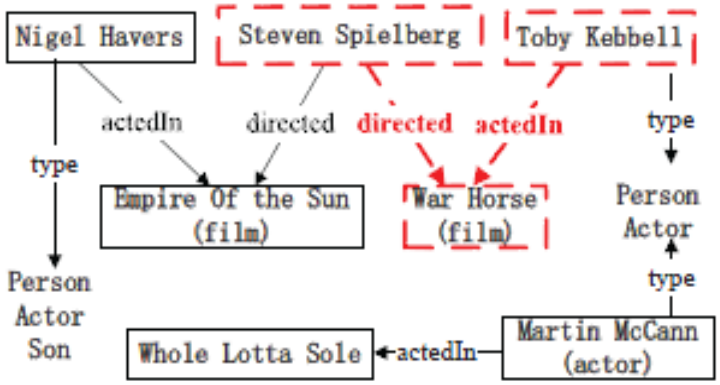
$$\text{Person} \xrightarrow{actedIn} \text{Movie} \xrightarrow{directed^{-1}} \text{Person} \xrightarrow{directed} \text{Movie} \xrightarrow{actedIn^{-1}} \text{Person}$$

  - Semantic meaning:

Actors of the movies directed by Steven Spielberg

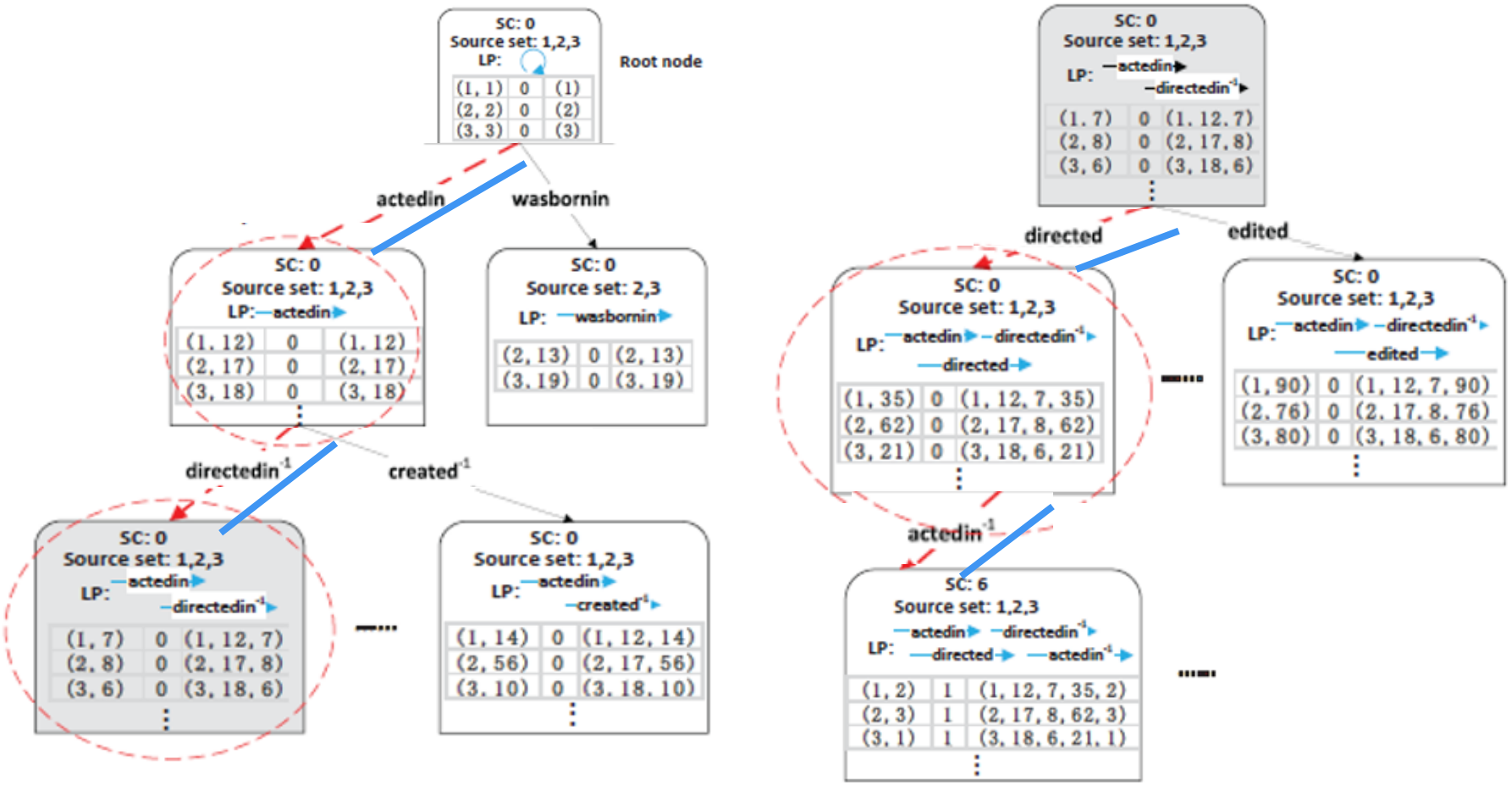
  - The expanded entity set:

{Jeremy Irvine, Teri Garr, Dee Wallace, Morgan Freeman…}



- Challenges
  - How to automatically generate meta paths
  - How to determine their weights

- Seed-based Meta Path Generation（SMPG）
  - Greedy tree that generate a relation sequence connecting more seeds with DFS.



- Predicting function

$$R(c_i, S) = \frac{1}{m} \sum_{j=1}^{m} \sum_{k=1}^{l} w_k \, r\{(c_i, s_j)|p_k\}$$

- Weight Learning
  - Heuristic method

the number of seed pairs connected by the meta path

$$w_k = \frac{\frac{|SP_k|}{m*(m-1)}}{\sum_{k=1}^{l} \frac{|SP_k|}{m*(m-1)}} = \frac{|SP_k|}{\sum_{k=1}^{l} |SP_k|}$$

  - PU learning method
    - The seed pairs are positive data, others are unlabeled data.

Experimental Setting

- Yago dataset
  - contain 35 relationships, more than1.3 million entities of 3455 instance classes.

Table 1. Description of the data.

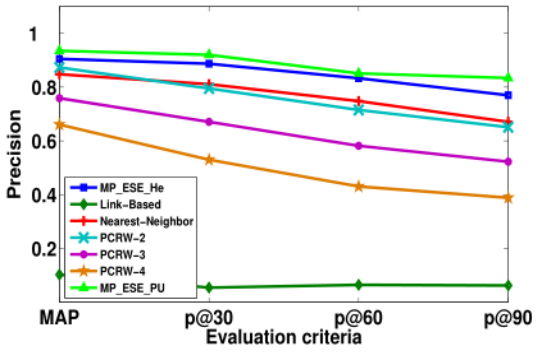
Data	Template of triples	# triples
yagoFacts	< entity relatinship entity >	4,484,914
yagoSimpleTypes	< entity rdf:type wordnet_type >	5,437,179
yagoTaxonomy	< wordnet_type rdfs:subclassof wordnet_type >	69,826

- Six tasks:
  - Actor, Software, Movie, Scientist, Writer and Married_actor

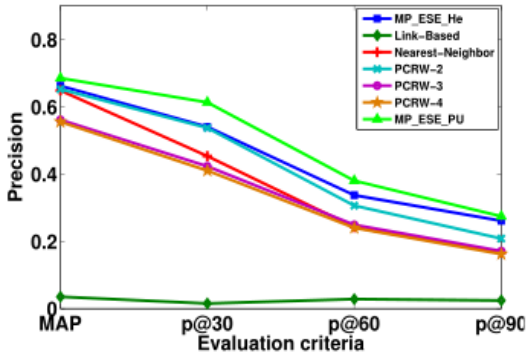
Actor: actors of the movies directed by Steven Spielberg,
Software: softwares of the companies located in Mountain View of California,
Movies: movies whose director won National Film Award,
Scientist: scientists of the universities located in Cambridge of Massachusetts,
Writer: writers being graduated from the universities in New York,
Married_actor**:** won Emmy Award and their spouses are also actors)

Effectiveness Experiments

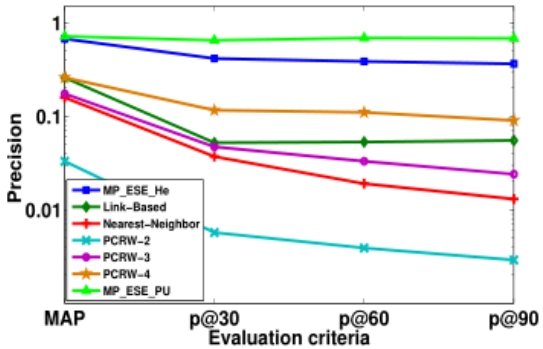
- Results of effectiveness experiments



(d) Scientist*



(e) Writer*



(f) Married_actor*

actors of the movies directed by Steven Spielberg

TABLE 3
Top 3 meta paths, heuristic and PU weights for Actor* task.

meta path	heuristic weight	PU weight
$\text{Person} \xrightarrow{actedIn} \text{Movie} \xrightarrow{directed^{-1}} \text{Person} \xrightarrow{directed} \text{Movie} \xrightarrow{actedIn^{-1}} \text{Person}$	0.2180	0.2082
$\text{Person} \xrightarrow{actedIn} \text{Movie} \xrightarrow{writeMusicFor^{-1}} \text{Person} \xrightarrow{writeMusicFor} \text{Movie} \xrightarrow{actedIn^{-1}} \text{Person}$	0.1495	0.1561
$\text{Person} \xrightarrow{actedIn} \text{Movie} \xrightarrow{edited^{-1}} \text{Person} \xrightarrow{edited} \text{Movie} \xrightarrow{actedIn^{-1}} \text{Person}$	0.1476	0.1399

Two people act in the movies directed by the same director