Mining relationships in Heterogeneous Academic Networks

Xiao Zeng 515030910531
Yinan He 515030910532

Outline for Section 1

Problem Description

Relationship identification

- Heterogeneous academic networks with papers, authors and venues
- Mining relationships by network structure & content information

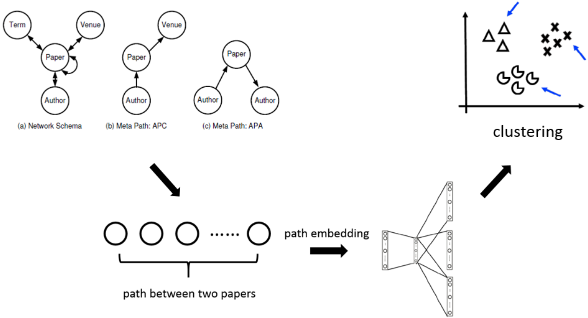
Figure: Problem Description

Outline for Section 2

Path2vec Model

Graph Structure Model Framework

- Generating heterogeneous node sequence
- Embedding
- Adaptive clustering



Path2vec Model

Path Pattern

We use the following metapath pattern to guide random walkers.

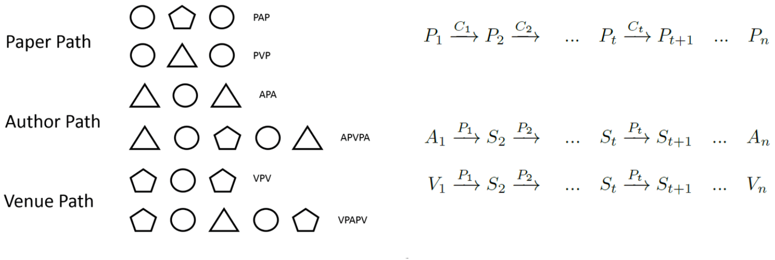


Figure: Path Pattern

Path2vec Model

Path Pattern

The transition probability is denoted as follows.

$$P(v^{i+1}|v_t^i, \mathscr{P}^{1}) = \begin{cases} \frac{1}{|N_{t+1}(v_t^i)|} & (v^{t+1}, v_t^i) \in E, \phi(v^{i+1} = t+1) \\ 0 & (v^{t+1}, v_t^i) \in E, \phi(v^{i+1} \neq t+1) \\ 0 & (v^{t+1}, v_t^i) \notin E \end{cases} \quad (1)$$

where $v_t^i \in V_t$ and $N_{t+1}(v_t^i)$ denote the V_{t+1} type neighborhood of node v_t^i.

[1]P refers to a meta-path scheme

Path2vec Model

Adaptive Clustering

- To classify the nodes into a rational number of categories
- Adjust the clustering number by setting a threshold δ

$$\delta = \beta \max_{v_i, v_j \in V} d(v_i, v_j)$$

- Get the least k that satisfy the threshold δ: optimal k

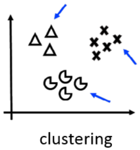


Figure: Adaptive Clustering

Experiments

Datasets

DBLP

Number of papers: 1.2M

Number of authors: 710K

Number of venues: 5K

DBIS

Number of papers: 10K

Number of authors: 5K

Number of venues: 424

AMiner

Number of papers: 30k

Number of authors: 90K

Number of venues: 3800+

Results

Embedding Visualization

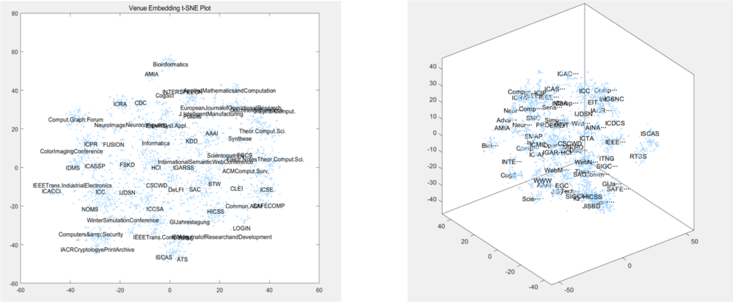


Figure: Venue Embedding Visualization

Results

Embedding Visualization

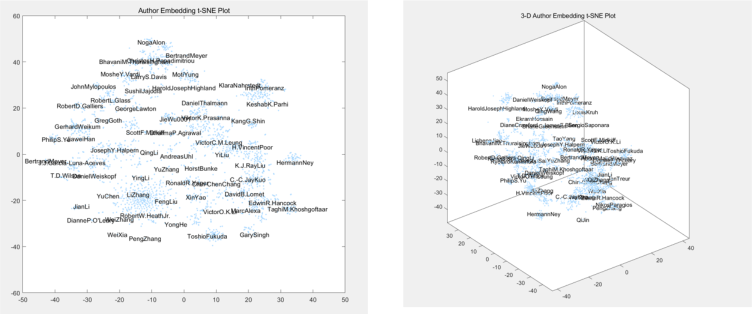
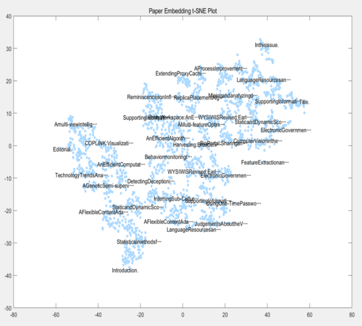


Figure: Author Embedding Visualization

Results

Embedding Visualization



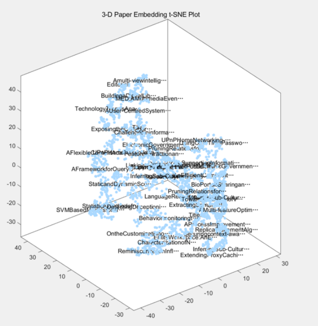


Figure: Paper Embedding Visualization

Results

Embedding Clustering Evaluation

- Compactness CP
- Davies-Bouldin Index DB(DBI)

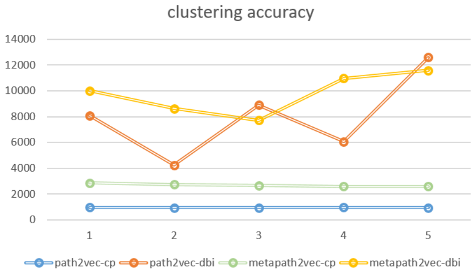


Figure: Embedding Clustering Evaluation

Results

Venue Embedding Clustering Visualization

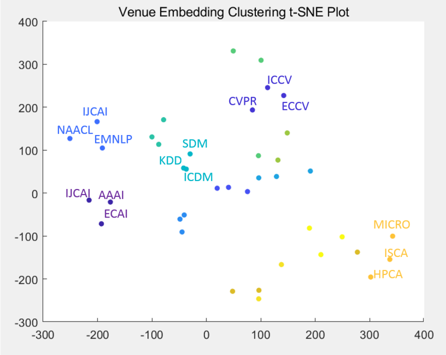


Figure: Venue Embedding Clustering Visualization

Case Study

Venue Similarity Search

Table: Case study of Venue similarity search in AMiner Data

ACL		NIPS		INFOCOM	
ACL	1	NIPS	1	INFOCOM	1
EMNLP	0.966946	ICML	0.955095	IEEE/ACM TN	0.980632
CL	0.959138	AISTATS	0.945436	MobiHoc	0.939416
CoNLL	0.933703	NC	0.908183	MobiCom	0.91177
IJCNLP	0.922411	COLT	0.89621	SECON	0.905895
COLING-ACL	0.914321	UAI	0.873626	IWQoS	0.904628
NLE	0.913332	CVPR	0.842136	GLOBECOM	0.896472
LREC	0.902107	KDD	0.84182	WiOpt	0.896011
EACL	0.900098	ACML	0.832118	vCoNEXT	0.890572
ANLP	0.899777	ECCV	0.830614	SIGCOMM	0.888044
LREC	0.888303	AAAI	0.824888	vICC	0.885082

Case Study

Author Similarity Search

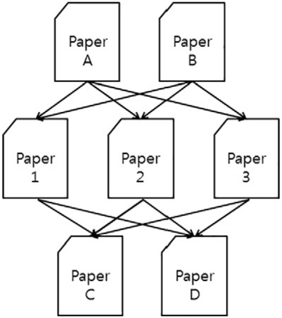
Table: Case study of Author similarity search in AMiner Data

LuoyiFu		JohnE.Hopcroft	
LuoyiFu	1	JohnE.Hopcroft	1
XinbingWang	0.887362	PrabhakarRaghavan	0.868889
WenyeWang	0.873246	AllanBorodin	0.842907
MichalisTitsias	0.864281	C.Seshadhri	0.829507
BenLiang	0.857967	AndrewChi-ChihYao	0.828785
KejieLu	0.857905	RudolphLanger	0.825836
aShiwenMao	0.856818	RobertEndreTarjan	0.825729
aShangqianHu	0.855743	RasmusPagh	0.82545
aKiTaekLee	0.852274	JurisHartmanis	0.821832
aMarcoFeletig	0.850694	JakubOcwieja	0.819203
aUlasC.Kozat	0.84784	VikrantSinghal	0.818596

Citation-based Model

Citation Network

- Directed graph $G = (V, E)$
- $V = \{v_i | i = 1, 2, \dots, n\}$
- $E = \{e_{ij} | i = 1, 2, \dots, n; j = 1, 2, \dots, n; i \neq j\}$



Citation-based Model

Measure Relevance

Katz Graph Distance Measure The relevance between two nodes x and y in a graph can be defined as:

$$R(x, y) = \Sigma_{p_i \in P} \eta^{|p_i|} \tag{2}$$

Where $\eta \in [0, 1]$ is a decay parameter, P denote the set of all paths between x and y.

Citation-based Model

Measure Relevance

- **Weighted Citation Link** $e_{ij} = (v_i, v_j, \omega_{ij})$, where ω_{ij} reflects the citation type.

$$\omega_{ij} = \begin{cases} \omega_0, \text{ paper j is an important citation of paper i} \\ \omega_1, \text{ paper j is an unimportant citation of paper i} \end{cases}$$

- **Weighted Katz Graph Distance Measure**

$$R(x, y) = \Sigma_{p_i \in P} \eta^{|p_i|} \frac{\Sigma_{e_{jk} \in p_i} \omega_{jk}}{|p_i|} \tag{3}$$

Citation Classification Model

Generating Weight

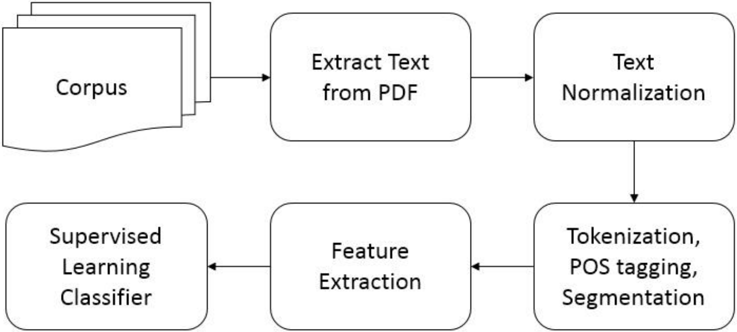


figure: model flowchart

Citation Classification Model

Features

Number of direct citation	Cue words similarity for important class
Number of direct citation/Number of all direct citation	Cue words similarity for unimportant class
Number of direct citation per section	Abstract similarity
Number of indirect citation	Number of paper that cited the citation per year
Number of indirect citation/Number of all indirect citation	Citation appears in table or caption
Number of indirect citation per section	

figure: List of Features

Citation Classification Model

Dataset for Training

Label	Number of Citation
Important	322
Unimportant	2811

Label	Number of Citation
Important	67
Unimportant	398

Label	Number of Citation
Important	389
Unimportant	3209

Label	Number of Citation
Important	389
Unimportant	673

figure: Annotated Dataset

Citation Classification Model

Training Result

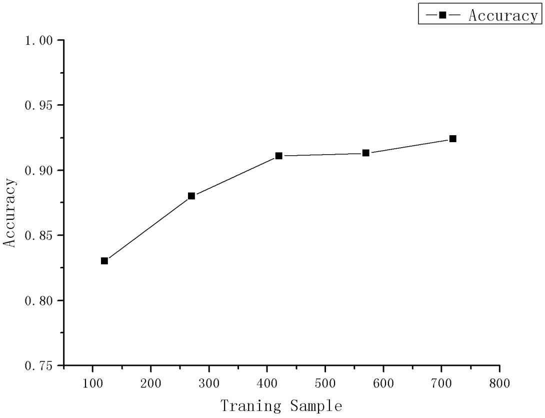


figure: Training Curve of SVM

Citation Classification Model

Training Result

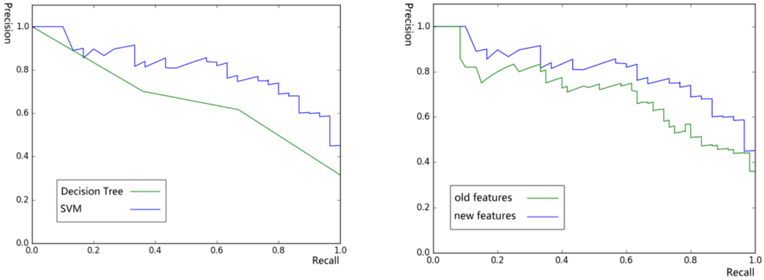


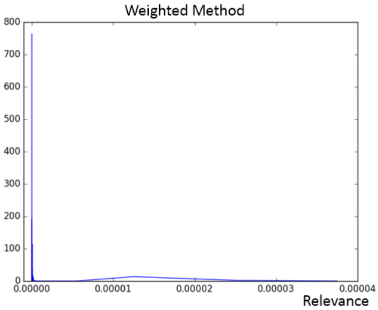
figure: Precision-Recall Curve Comparison

Citation-based Model

Experiment Result

Paper	Citation	Author	Venue
21290	12382	14981	341

figure: Dataset



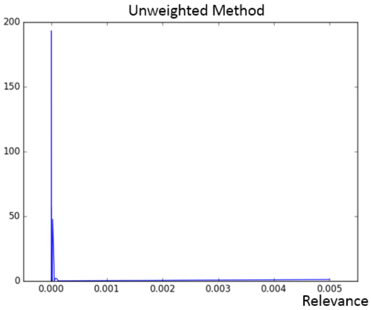


figure: Relevance Distribution

Outline for Section 3

Future work

- Integrate path2vec on heterogeneous academic networks and citation-based method on citation network
- Increase efficiency in classifying citations
- Incorporate path bias in applying path generation strategy

Outline for Section 4

Bibliography

1. Yuxiao Dong, Nitesh V. Chawla, Ananthram Swami: metapath2vec: Scalable Representation Learning for Heterogeneous Networks. KDD 2017: 135-144
2. Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, Tianyi Wu: PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. PVLDB 4(11): 992-1003 (2011)
3. Liben-Nowell, D., Kleinberg, J.: The Link-prediction Problem for Social Networks. Journal of the American Society for Information Science and Technology 58(7), 1019–1031 (2007)
4. lenzuela M, Ha V, Etzioni O.Identifying Meaningful Citations AAAIWorkshop: Scholarly Big Data.2015.
5. Zhu, X.; Turney, P.; Lemire, D.; and Vellino, A. 2013. Measuring academic influence: Not all citations are equal. submitted to Journal of the Association for Information Science and Technology (JASIST).

THANK YOU!