论文作者名消歧
(Author Name Disambiguation)

郭晨亮

task

- 作者重名现象，同一名字对应多个作者、多篇论文，需要区分。
- 应用：信息检索、文献计量、学术知识图谱构建
- 歧义：同一名字的不同表示(词序、缩写、错字)

  多个人具有同一名字(名字本身相同、缩写/出错后相同)

- 论文相关信息：

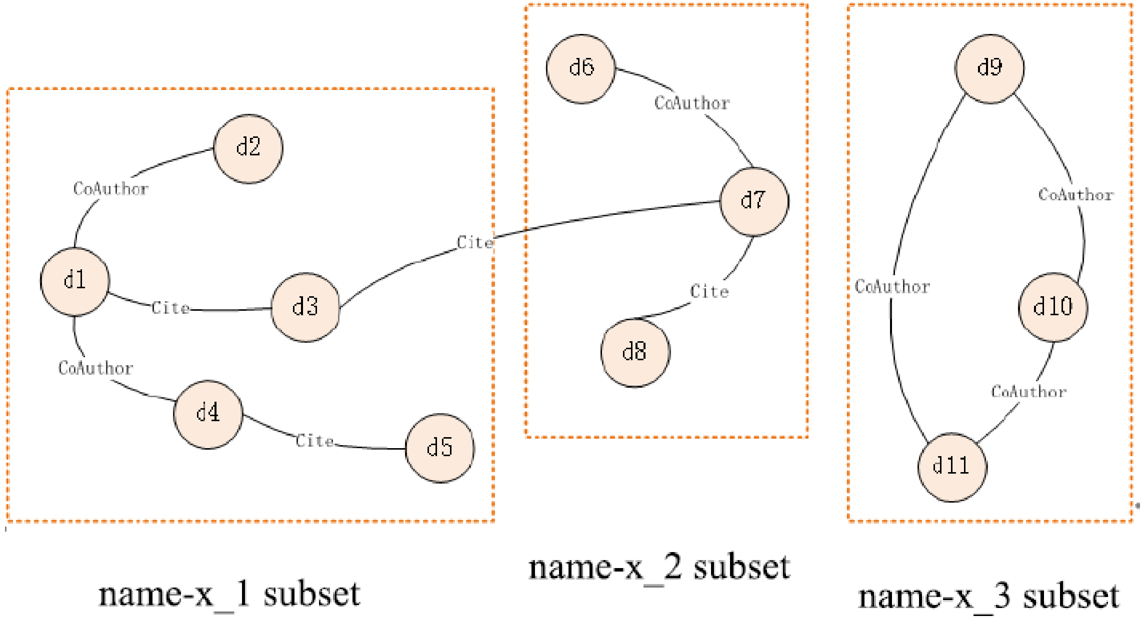
标题(title)、摘要(abstract)、多个作者(authors)、机构(organization)、出版机构(venue)、年份(year)、关键词(keyword)

作者个人主页、邮箱、地址

引用关系

- 方法：对于一组同名作者论文集，构建关系、学习论文特征表示、聚类。

task



- {P1,P2}, {P3}, {P4,P5}

ID	关键词	机构	其它作者列表
P1	adaptive computing allocation mobile cloud	School of Transportation and Logistics	tianyi xing;lin x cai;dijiang huang;daiyuan peng;yan liu
P2	adaptive channel allocation wireless algorithm	National United Engineering Laboratory of Integrated and Intelligent Transportation	jin zhang;wei li;t aaron gulliver
P3	network coding DVB-IPDC LTE	Secure Networking and Computing Institute, Arizona State University	lian wang;daiyuan peng
P4	5-axis machining G code Interpolation NURBS	College of Mechanical Engineering	xia li
P5	5-axis NURBS surfaces STEP-NC	College of Mechanical and Electrical Engineering	xia li

作者hongbin liang的论文数据

task

分类：

- 冷启动：从已有的论文数据库中得到一个消歧初始结果
- 增量更新：随着时间变化论文增加将新的歧义作者名补充到已有消歧结果

现有问题:

- 1.如何有效利用全局和局部信息
- 2.如何融合不同类型、不同影响程度的异构特征
- 3.如何融合文本、结构关系
- 4.减少信息缺失的影响
- 5.如何在不知道同一人名对应人数的情况下正确聚类
- 6.利用已知标记信息的方式
- 7.合理设置融合参数、不同语言的影响

- DBLP(679,2.2) Aminer(110,13.8/100,4.9) CiteSeerX (14,33.4)

task

方法分类：

- 监督学习：SVM、分类任务（不适用于缺少标注、大量的数据）
- 无监督学习：

(1) 图表示学习方法：LINE、Node2vec、DeepWalk、HIN2vec、GNN等

(2) 聚类方法：层次凝聚聚类(HAC)、层次聚类、DBSCAN等

- 互联网资源、人工参与标记
- 聚类数量K的预测
- 对抗学习

评价指标：

属于同一类且分类为同一类的论文对数量称为真阳性TP

属于同一类且分类为不同类的论文对数量称为假阴性FN

属于不同类且分类为同一类的论文对数量称为假阳性FP

召回率Recall　　精确率Precision

$$Recall = \frac{TP}{TP+FN},\ Precision = \frac{TP}{TP+FP},$$

$$F1_i = \frac{2}{\frac{1}{prec_i} + \frac{1}{recall_i}} = \frac{2 \cdot prec_i \cdot recall_i}{prec_i + recall_i} = \frac{2n_{ij_i}}{n_i + m_{j_i}}.$$

$$Precision = \frac{1}{r}\sum_{i=1}^{r} prec_i \quad Recall = \frac{1}{r}\sum_{i=1}^{r} recall_i. \quad F1 = \frac{1}{r}\sum_{i=1}^{r} F1_i.$$

Diting

$$s(v_i, v_j) = \frac{d_i^T d_j}{\|d_i\| \cdot \|d_j\|}$$

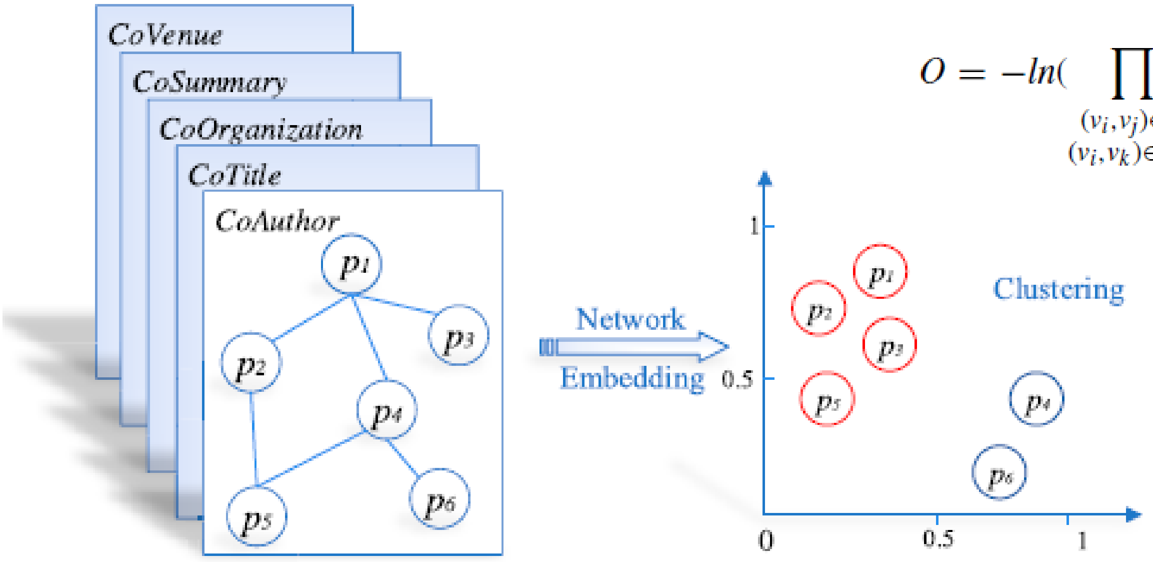
$$L(i,j,k) = max(\epsilon, s(v_i, v_j) - s(v_i, v_k))$$

$$O = -ln(\prod_{\substack{(v_i,v_j)\in E \\ (v_i,v_k)\in NE}} L(i,j,k))$$

$$O_{\mathcal{N}} = \sum_{\substack{(v_i,v_j)\in E^{\mathcal{N}} \\ (v_i,v_k)\in NE^{\mathcal{N}}}} -ln(max(\epsilon, s(v_i, v_j) - s(v_i, v_k)))$$

$$\mathcal{N} \in \{a, t, v, s, o, y\}$$

$$O_{Diting} = \sum_{i\in\{a,t,v,s,o,y\}} w_i O_i + \lambda L^2$$



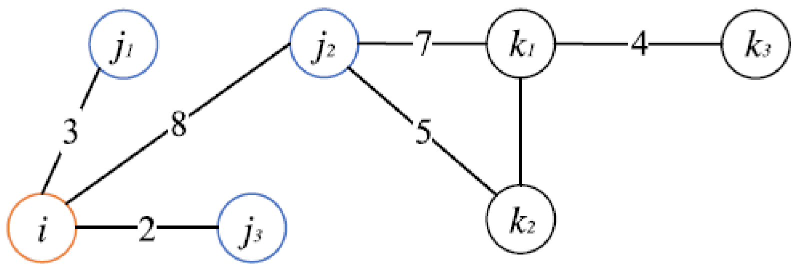
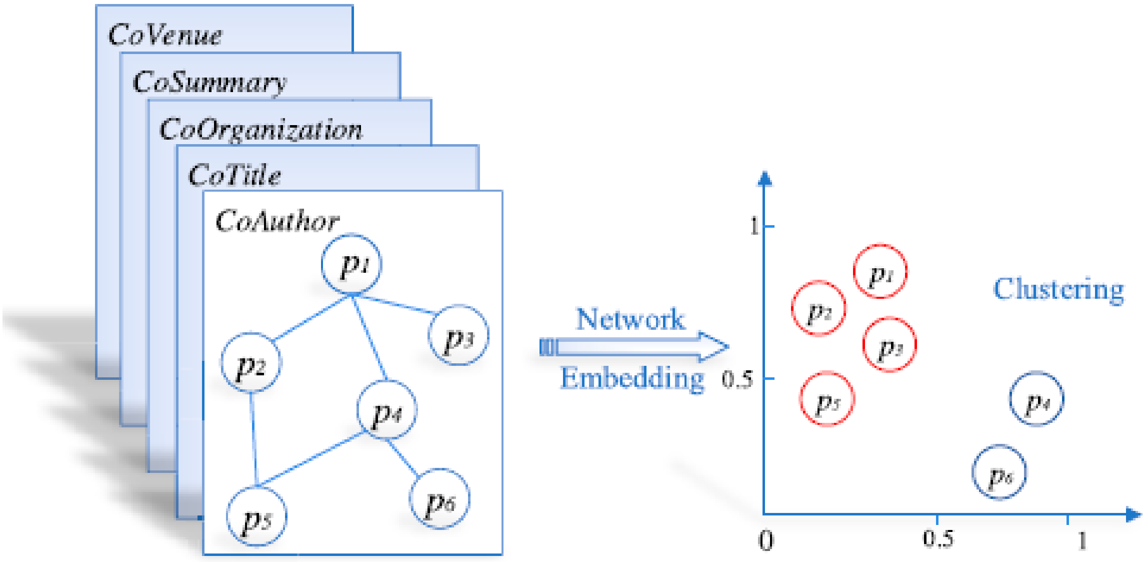


FIGURE 2. An example of positive and negative edge sampling.

- 作者、标题、摘要、出版机构、机构、年份、标签
- Coauthor:对作者名缩写形式、不同顺序进行处理
- Cotitle:用NLTK做词性还原，用NTEE计算标题词嵌入，若相似度过阈值为相同。
- CoSummary:抽取固定数量的关键词
- CoVenue:缩写转全称

- Diting: An Author Disambiguation Method Based on Network Representation Learning, IEEE 2019

Diting



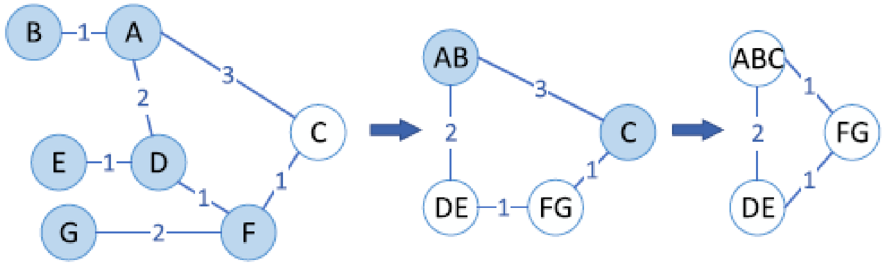


FIGURE 3. Illustration of the coarsening procedure (shaded neighboring nodes are merged).

$$SD = \max_{i,j} d(c_i, c_j) scat + dis$$

$$scat = \frac{1}{M}\sum_i \|\sigma(c_i)\| / \|\sigma(D)\|$$

$$dis = \frac{\max_{i,j} d(c_i, c_j)}{\min_{i,j} d(c_i, c_j)} \sum_i \Big(\sum_j d(c_i, c_j)\Big)^{-1}$$

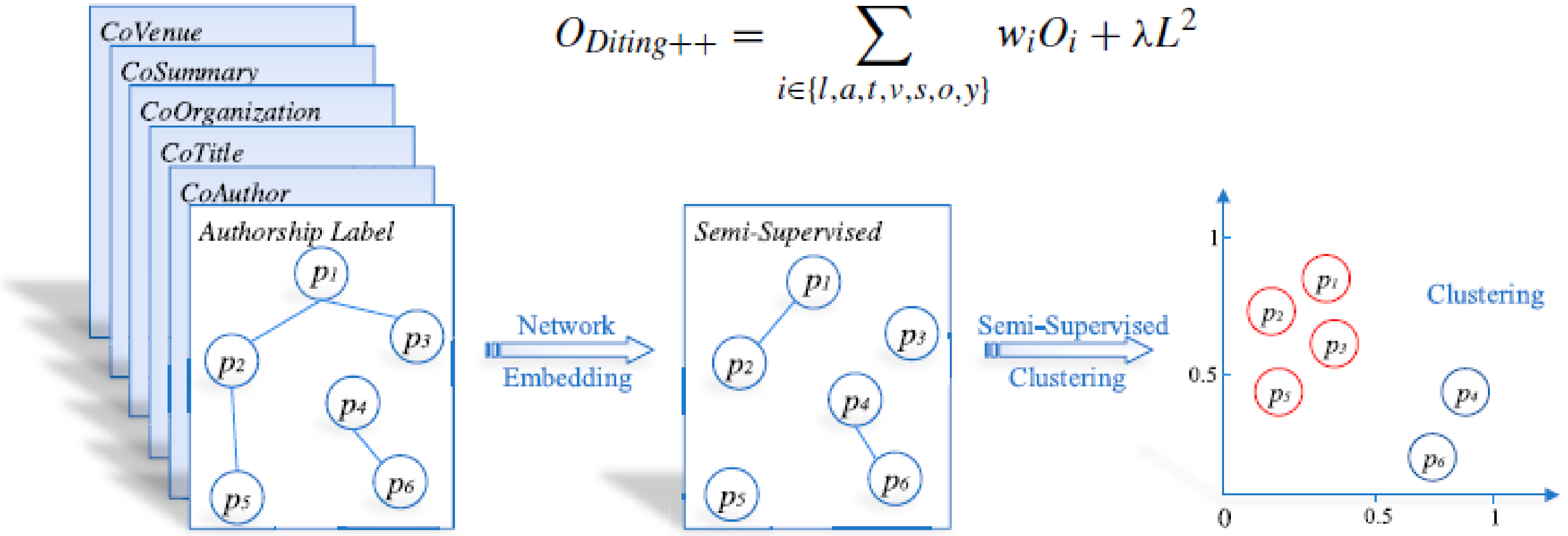
$$\sigma(c_i) = \frac{1}{n_i}\sum_j d(p_j, c_i)^2.$$

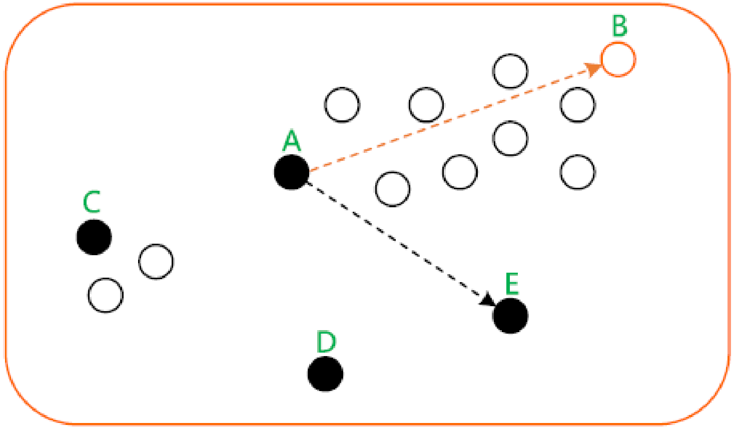
- 网络粗化到1/3
- 对于倾斜数据：K-means、AP、DBSCAN
- K-means需要聚类数，其它两种容易过度合并孤立点，结合AP和DBSCAN的HDBSCAN
- SD第一项表示聚类内方差大小，第二项表示不同聚类间的距离

- Diting: An Author Disambiguation Method Based on Network Representation Learning, IEEE 2019

Diting++

$$O_{Diting++} = \sum_{i\in\{l,a,t,v,s,o,y\}} w_i O_i + \lambda L^2$$





$$p(v_i, c_{i-1}) = \frac{d(v_i, c_{i-1})^2}{\sum_{v'\in\mathcal{P}^n} d(v', c_{i-1})^2}$$

$$d(v_i, v_j) = \frac{1}{2} + \frac{d_i \cdot d_j}{2\cdot\|d_i\|\cdot\|d_j\|}.$$

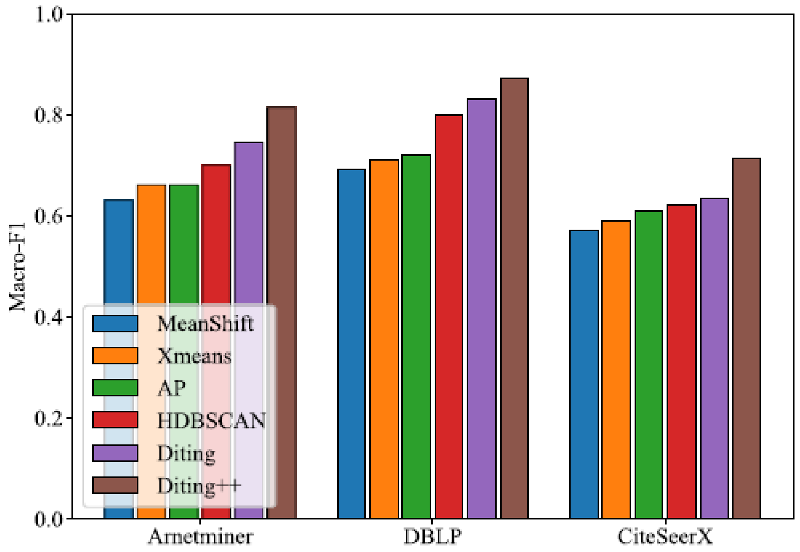
$$q(v_i, c_{i-1}) = \frac{1}{2}\cdot\frac{d(v_i, c_{i-1})}{\sum_{v'\in\mathcal{P}^n} d(v', c_{i-1})} + \frac{1}{2}\cdot\frac{\sum_{j=1}^{min(5,|\mathcal{P}^n|)} D_j^{v_i}}{min(5, |\mathcal{P}^n|)}$$

- Lable:两篇论文是否属于同一作者
- 用K-means聚类，初始点在一个类中最多选一个，选择聚类中心按距离越大概率越大随机。
- 受到COP-Kmeans的启发，为了避免k-means导致将一个大类拆分，增加top5与最近点均距离提升孤立点被选为聚类中心的概率，并将约束用于分配过程。

- Diting: An Author Disambiguation Method Based on Network Representation Learning, IEEE 2019

Diting++

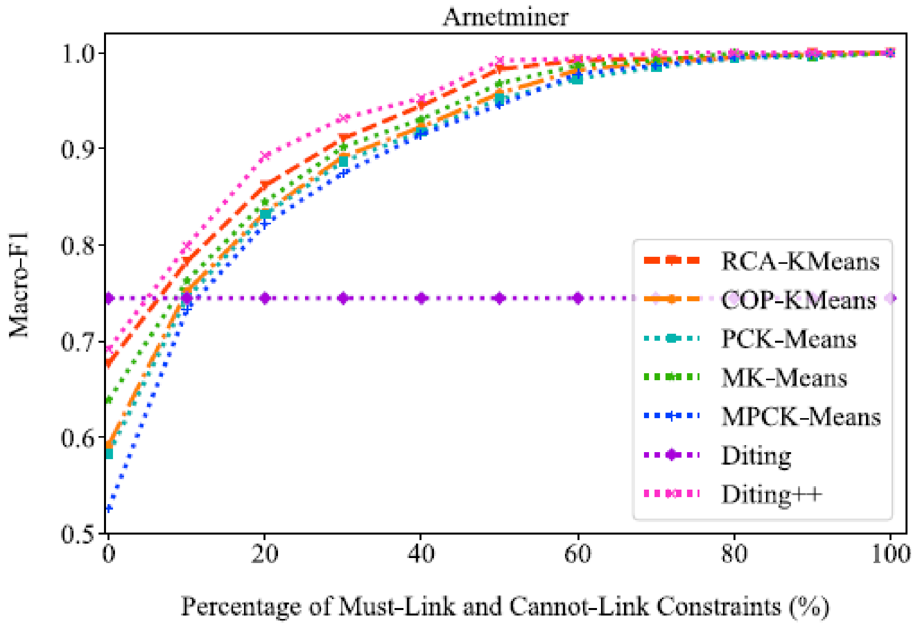
	Arnetminer			DBLP			CiteSeerX		
Method	P	R	F	P	R	F	P	R	F
Khabsa *et al.* 2015 [42]	0.633	0.571	0.584	0.674	0.626	0.645	0.421	0.447	0.424
Qian *et al.* 2015 [43]	0.604	0.512	0.547	0.686	0.656	0.681	0.553	0.521	0.547
Zhang *et al.* 2016 [11]	0.523	0.745	0.613	0.741	0.716	0.723	0.547	0.522	0.536
Zhang *et al.* 2017 [3]	0.576	0.693	0.635	0.758	0.706	0.742	0.597	0.575	0.596
DeepWalk 2014 [17]	0.621	0.558	0.582	0.724	0.756	0.734	0.477	0.513	0.482
LINE 2015 [19]	0.654	0.573	0.609	0.723	0.735	0.722	0.536	0.591	0.553
Node2Vec 2016 [22]	0.625	0.541	0.589	0.675	0.705	0.685	0.524	0.465	0.498
PTE 2015 [32]	0.697	0.568	0.632	0.741	0.794	0.762	0.548	0.611	0.578
CANE 2017 [44]	0.588	0.674	0.624	0.682	0.764	0.712	0.499	0.543	0.511
Hin2Vec 2017 [45]	0.655	0.561	0.616	0.714	0.755	0.743	0.589	0.517	0.562
Diting	**0.786**	**0.718**	**0.745**	**0.822**	**0.854**	**0.832**	**0.664**	**0.601**	**0.635**
Diting++	**0.853**	**0.738**	**0.814**	**0.846**	**0.896**	**0.871**	**0.744**	**0.684**	**0.712**

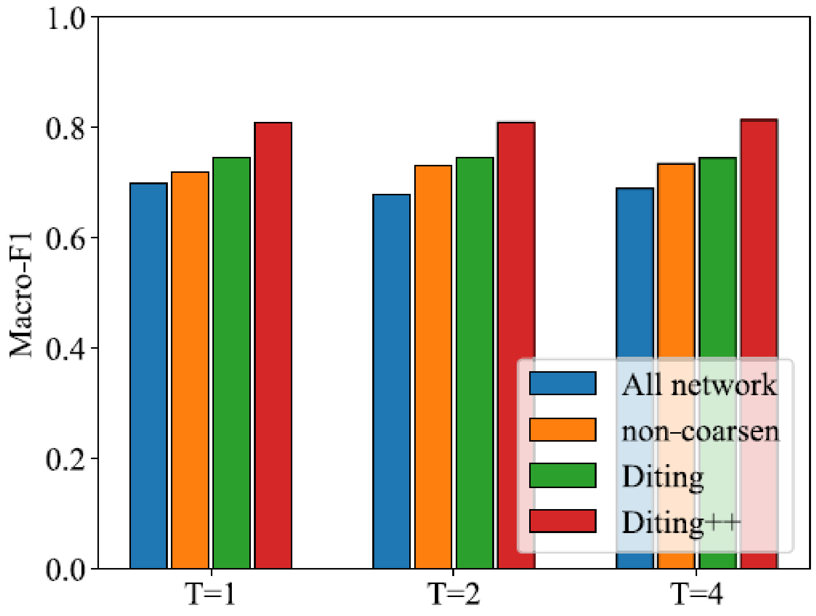


Name	Khabsa	Qian	Zhang16	Zhang17	DeepWalk	LINE	Node2Vec	PTE	CANE	Hin2Vec	Diting
Alok Gupta	0.564	0.571	0.672	0.652	0.618	0.625	0.636	0.745	0.681	0.566	**0.985**
Bin Li	0.615	0.591	0.682	0.676	0.545	0.558	0.579	0.486	0.608	0.582	**0.769**
Bing Liu	0.616	0.644	0.715	0.769	0.742	0.744	0.803	0.701	0.649	0.655	**0.982**
David Jensen	0.589	0.640	0.693	0.802	0.782	0.807	0.926	**0.932**	0.559	0.687	0.700
David Nelson	0.501	0.599	0.580	0.569	0.537	0.575	0.600	0.500	0.535	0.649	**0.785**
F. Wang	0.467	0.711	0.778	0.761	0.596	0.636	0.612	0.652	0.587	0.571	**0.912**
Jeffrey Parsons	0.771	0.785	0.722	0.768	0.655	0.723	0.744	0.824	0.601	0.533	**0.903**
Ji Zhang	0.492	0.491	0.486	0.513	0.496	0.646	0.492	0.521	0.735	0.638	**0.855**
Jie Yu	0.631	0.698	0.717	0.558	0.713	0.724	0.799	**0.831**	0.558	0.574	0.825
Jim Gray	0.644	0.675	0.789	0.754	0.681	0.832	0.863	0.942	0.613	0.711	**0.966**
Avg Macro-F1	0.591	0.592	0.640	0.651	0.593	0.645	0.631	0.668	0.616	0.597	**0.862**

EE 2019

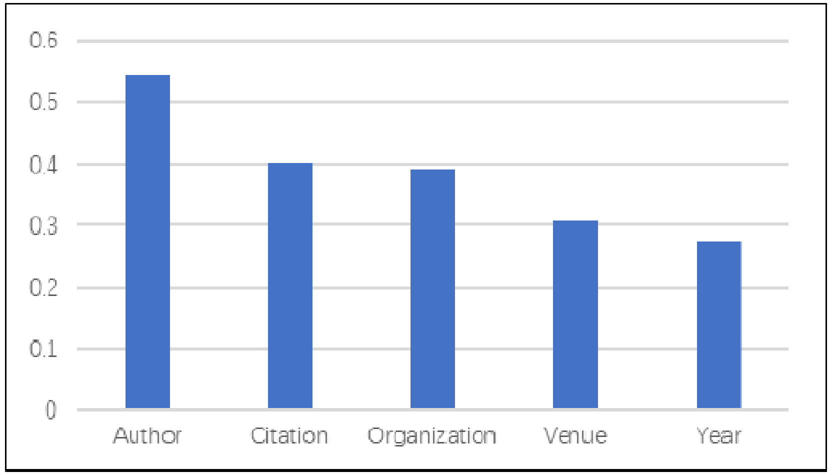
Diting++





- Diting: An Author Disambiguation Method Based on Network Representation Learning, IEEE 2019

HRFAENE



IGURE 4. Feature strength test results.

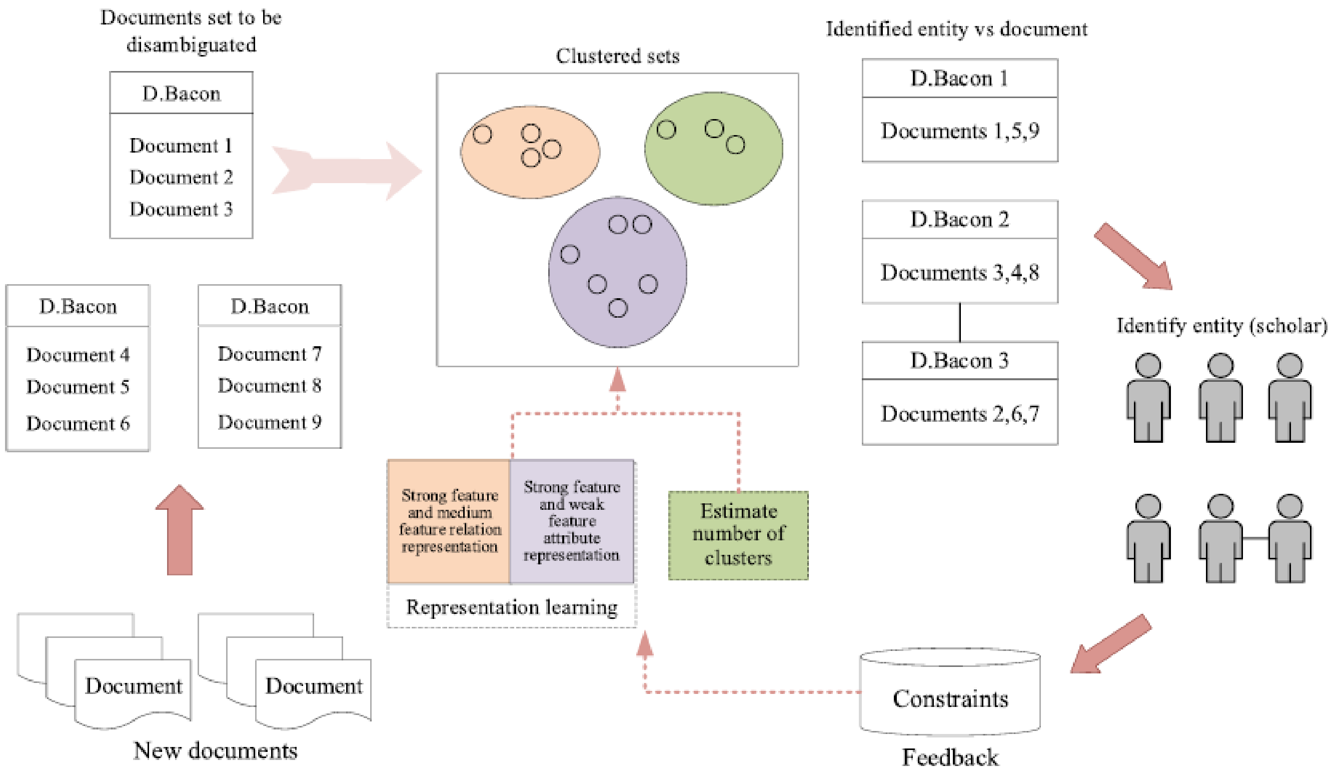
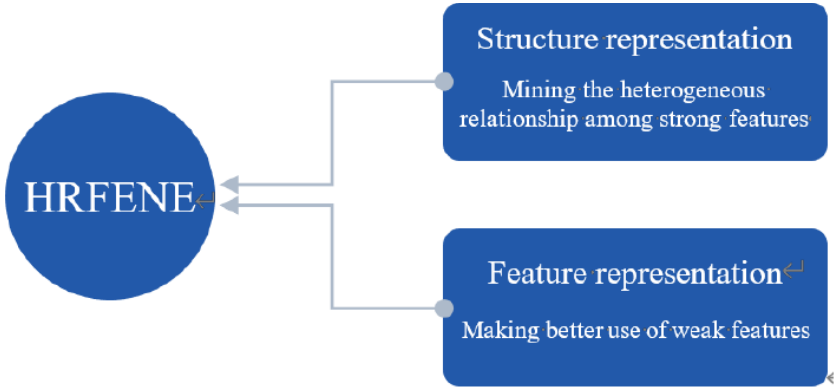


FIGURE 3. The overall framework of the name disambiguation.

- 不同特征影响不同，0.35,0.3分为强中弱特征
- （合作网络）某同名论文集中，作者名间的网络，边权为合著论文数
- （作者-文档网络）作者与论文关系，边权1
- （两步合作网络）对于两篇论文，论文作者的合著者的交集表示相似度
- （文档-特征网络）强特征与论文的关系，边权为特征次数



- Scholar Disambiguation Method Based on Heterogeneous Relation-Fusion and Attribute Enhancement, IEEE, 2020

HRFAENE

$$S^{\pi}_{v_i,v_j,v_t} = S(v_i, v_j) - S(v_i, v_t)$$

$$L(i,j,t) = \sum_{\substack{e_{ij}\in E^R \\ ne_{it}\in NE^R}} P\left(S(v_i, v_j) > S(v_i, v_t) \mid v_i, v_j, v_t\right)$$

$$\mathrm{R} \in \{aa, ad, dd, df\}$$

$$OBJ_s = \min_{\mathbf{A},\mathbf{D},\mathbf{F}} \sum_{R} \lambda_R O_R + \mathrm{RT}, \quad \mathrm{R} \in \{aa, ad, df, dd\}$$

$$P(\mathbf{f}_m|\mathbf{v}_i) = \frac{\exp(\mathbf{y}_i \cdot \mathbf{w}_m)}{\sum_{z=1}^{|\mathcal{F}|} \exp(\mathbf{y}_i \cdot \mathbf{w}_z)}$$

$$OBJ_f = -\min \sum_{i=1}^{|\mathcal{V}|} \sum_{m=1}^{|\mathcal{F}|} \theta_{im} \mathbf{X}_{im} \log P(\mathbf{f}_m|\mathbf{v}_i)$$

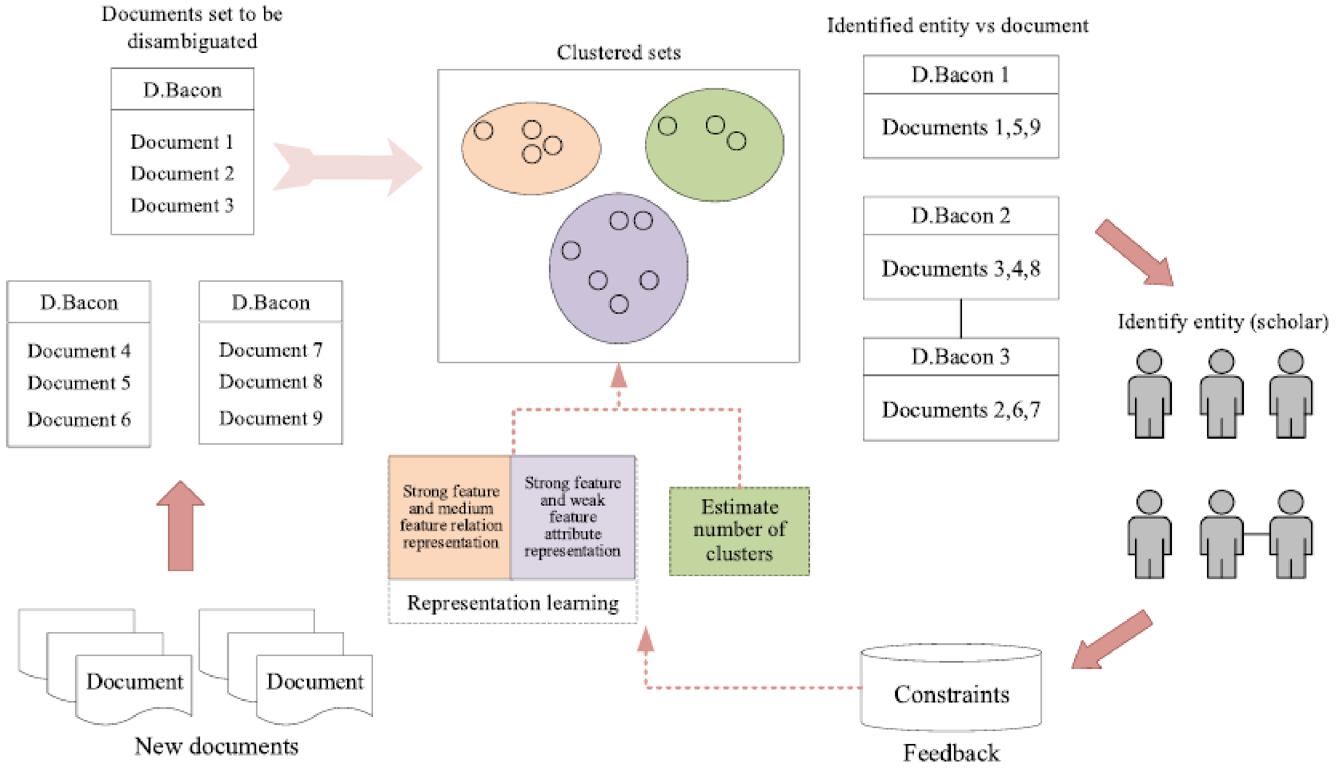


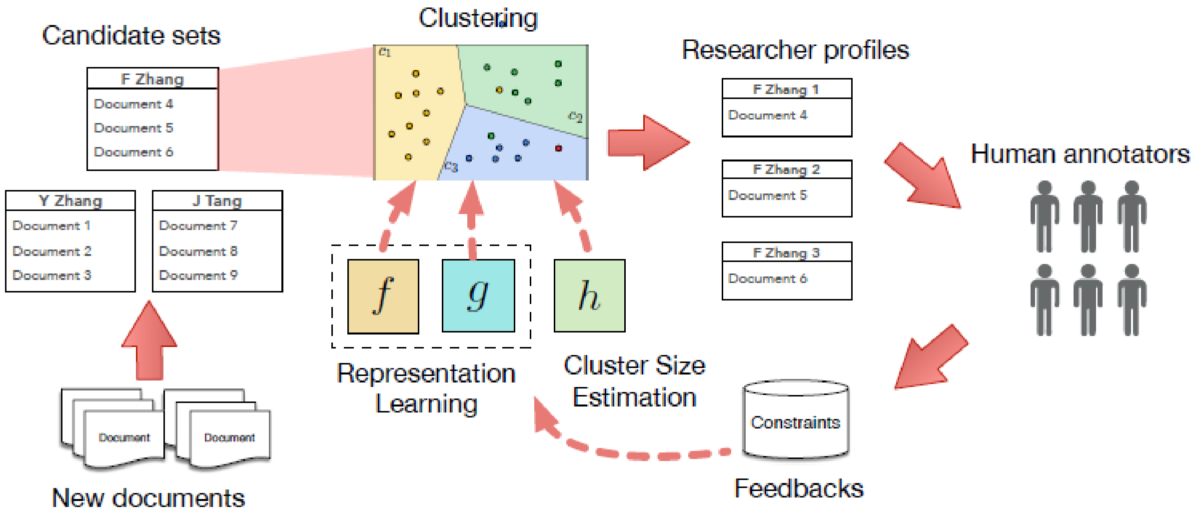
FIGURE 3. The overall framework of the name disambiguation.

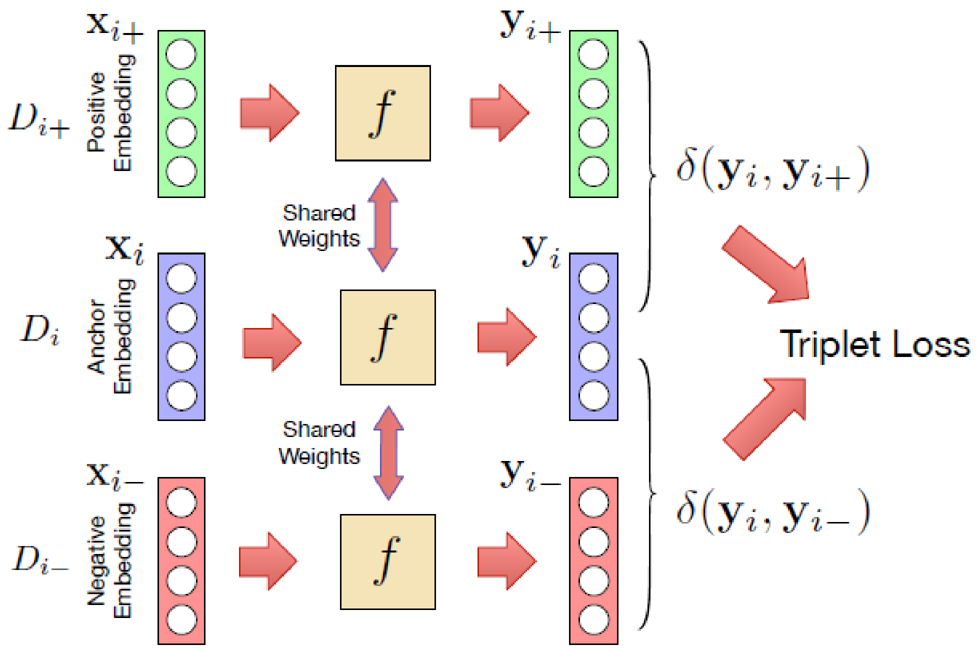
$$OBJ = \alpha \times OBJ_s + \beta \times OBJ_f$$

- 表示：成对约束可扩展的方法
- 取(vi,vj)相邻，(vi,vt)不相邻，正负样例相似度差sigmoid表示正大于负的概率，相似度为向量点乘，超参数加权求和+正则化
- 聚类：输入论文特征，经过线性层和softmax，取-log后求和最小化
- 测试：AMiner0.779

  - Scholar Disambiguation Method Based on Heterogeneous Relation-Fusion and Attribute Enhancement, IEEE, 2020

AMiner





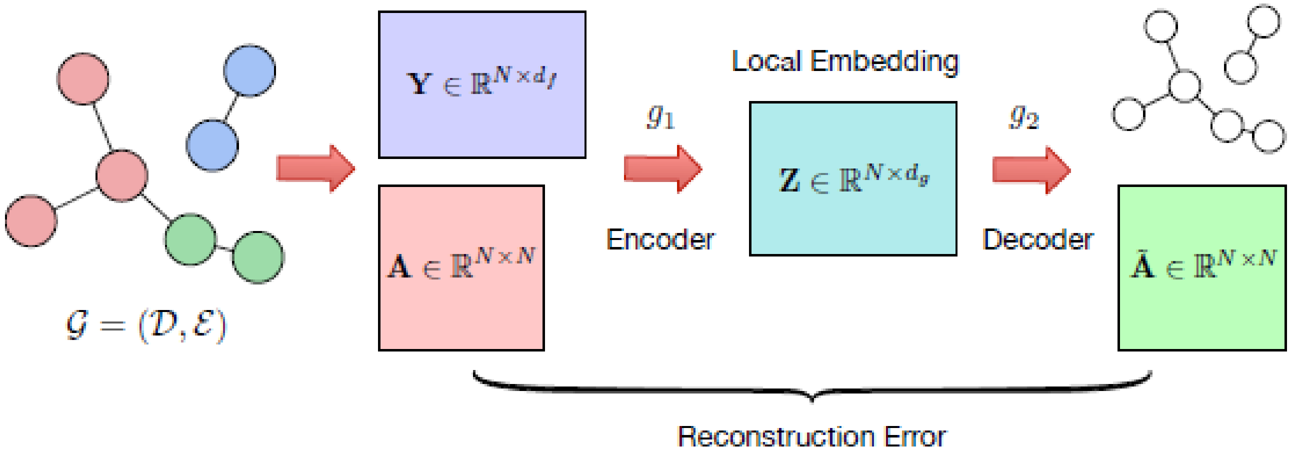
$$\begin{cases}(D_i, C_k, 0) \in S^I \rightarrow D_i \notin C_k, \\ (D_i, C_k, 1) \in S^I \rightarrow D_i \in C_k.\end{cases} \quad \begin{cases}(D_i, D_j, 0) \in S^P \rightarrow \mathbb{I}(D_i) \neq \mathbb{I}(D_j), \\ (D_i, D_j, 1) \in S^P \rightarrow \mathbb{I}(D_i) = \mathbb{I}(D_j).\end{cases}$$

$$\mathcal{L}_f = \sum_{(D_i, D_{i+}, D_{i-}) \in \mathcal{T}} \max\{0, \delta(\mathbf{y}_i, \mathbf{y}_{i+}) - \delta(\mathbf{y}_i, \mathbf{y}_{i-}) + m\}.$$

- 首先word2vec用IDF加权求和学习论文的全局表示，用约束正负3元组训练
- 用局部上下文微调，若论文共同特征IDF和大于阈值在局部图中连接边。
- 加入4中人工标注

- Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop, SIGKDD,2018

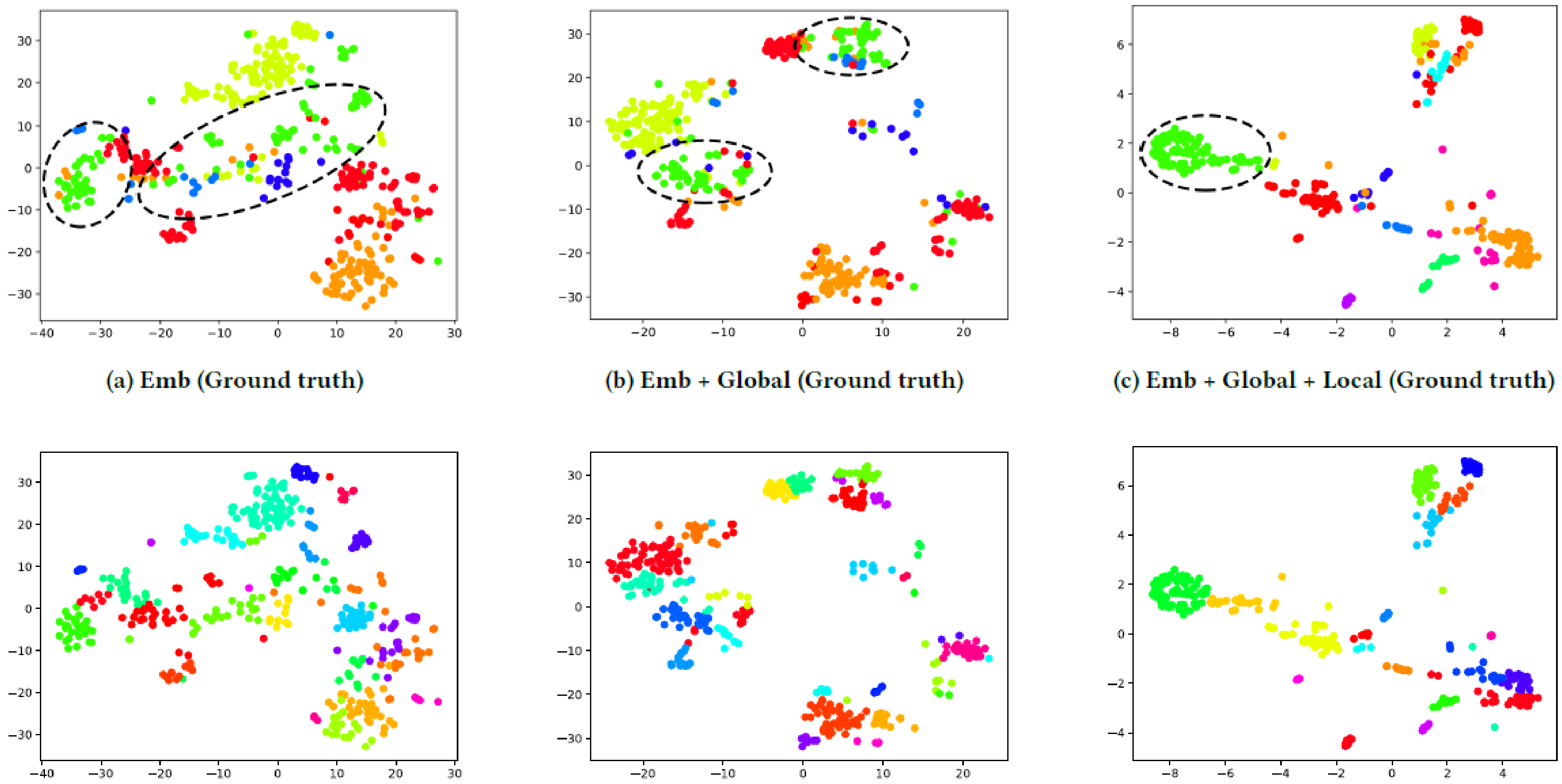
AMiner



- 用两层GCN的g1输入点表示和边做encoder，用g2层sigmoid点乘g1的输出预测边矩阵，最小化误差用于微调向量表示，用中间层得到新的论文表示。
- 用HAC聚类并估计聚类大小，以前的X-means方法有 1.倾向于过度聚类 2.无法适用于大量数据 的问题，用RNN输入一组论文向量，直接输出聚类数，用均方对数误差训练。
- 实时更新：用KNN分类器，全局更新索引匹配。
- 人参与：删除关联、增加关联、切分、合并、创建、确认

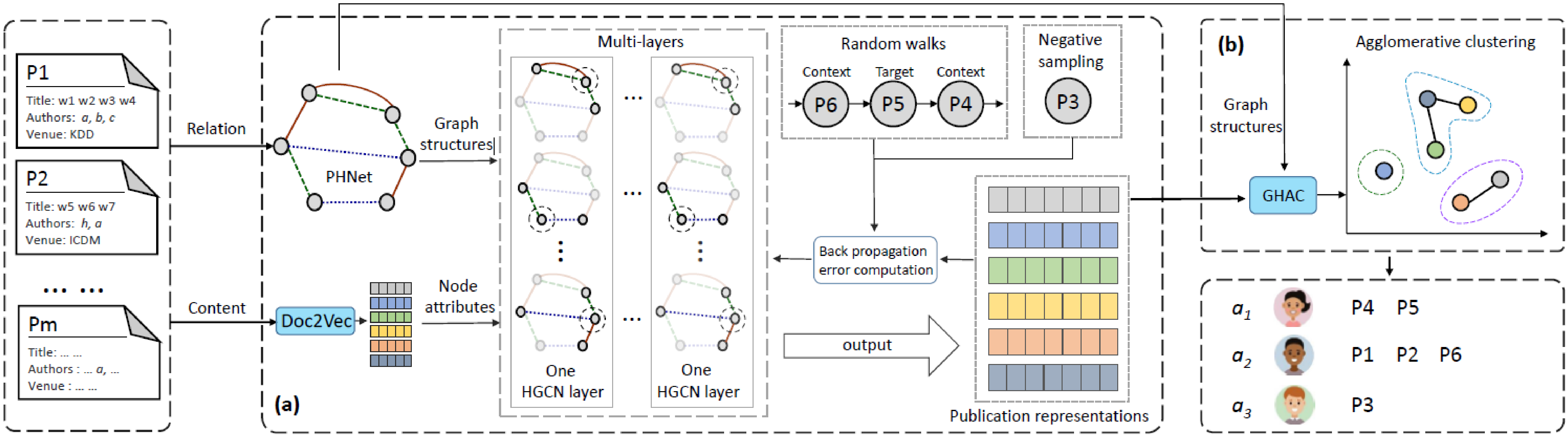
- Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop, SIGKDD,2018

AMiner


(a) Emb (Ground truth)
(b) Emb + Global (Ground truth)
(c) Emb + Global + Local (Ground truth)
(d) Emb (F1: 35.36%)
(e) Emb + Global (F1: 42.75%)
(f) Emb + Global + Local (F1: 61.11%)

- Name Disambiguation in AMiner: Clustering, Maintenance, and Human in the Loop, SIGKDD, 2018

HGCN



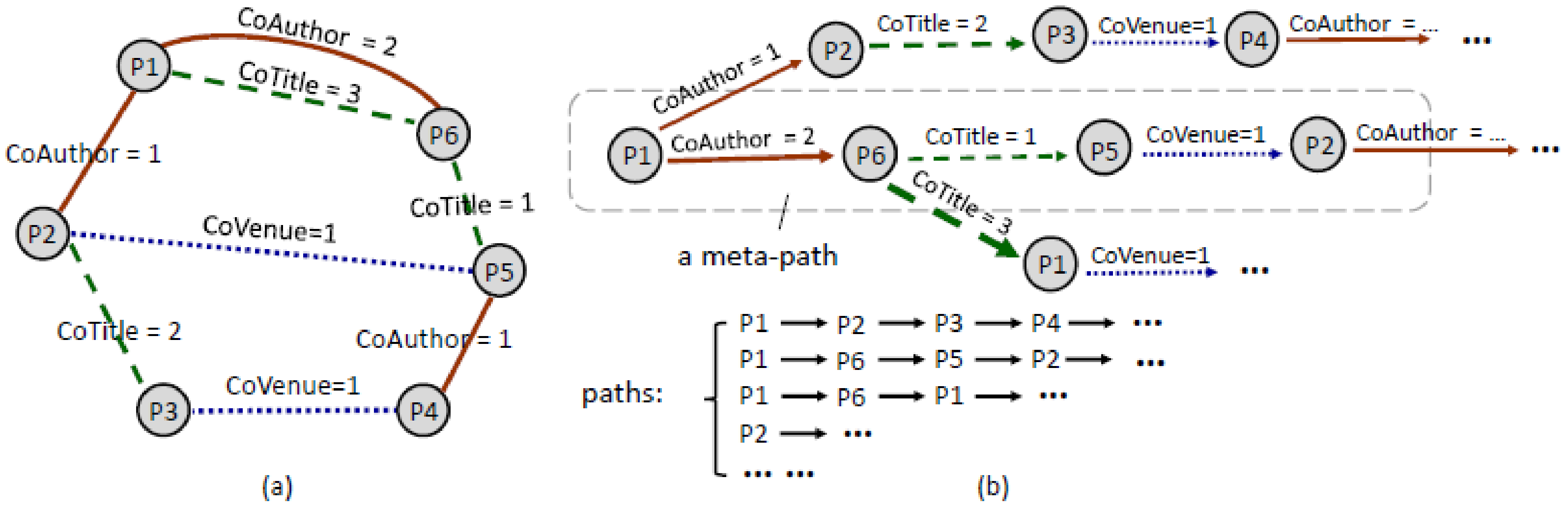
$$u_i^{(l+1)} = ReLU\left(\sum_{r\in\mathcal{R}}\sum_{j\in N_i^r}\frac{1}{c_{ij}^r}u_j^{(l)}W_r^{(l)}\right)$$

$$M = \frac{1}{2m}\sum_{p_i,p_j\in V}\left[|e_{ij}| - \frac{w_i w_j}{2m}\right]\delta(c_i = c_j)$$

- 同作者（数量）、同出版机构（1）、同标题（单词交集）构建论文间异构网络
- 用Doc2vec编码标题+摘要，2层HGCN，用关系对应不同权重矩阵传播
- 文本属性叠加结构属性，mini-batch Adam训练
- HAC擅长偏斜数据，用GHAC聚类，不需要预先的参数
- GHAC：改为同构图上的相邻距离，边权表示相似度，分组内边-分组外边最大化求K

- Unsupervised Author Disambiguation using Heterogeneous Graph Convolutional Network Embedding, IEEE,2019

HGCN



- Unsupervised Author Disambiguation using Heterogeneous Graph Convolutional Network Embedding, IEEE,2019

Using Graph Node Embedding Method

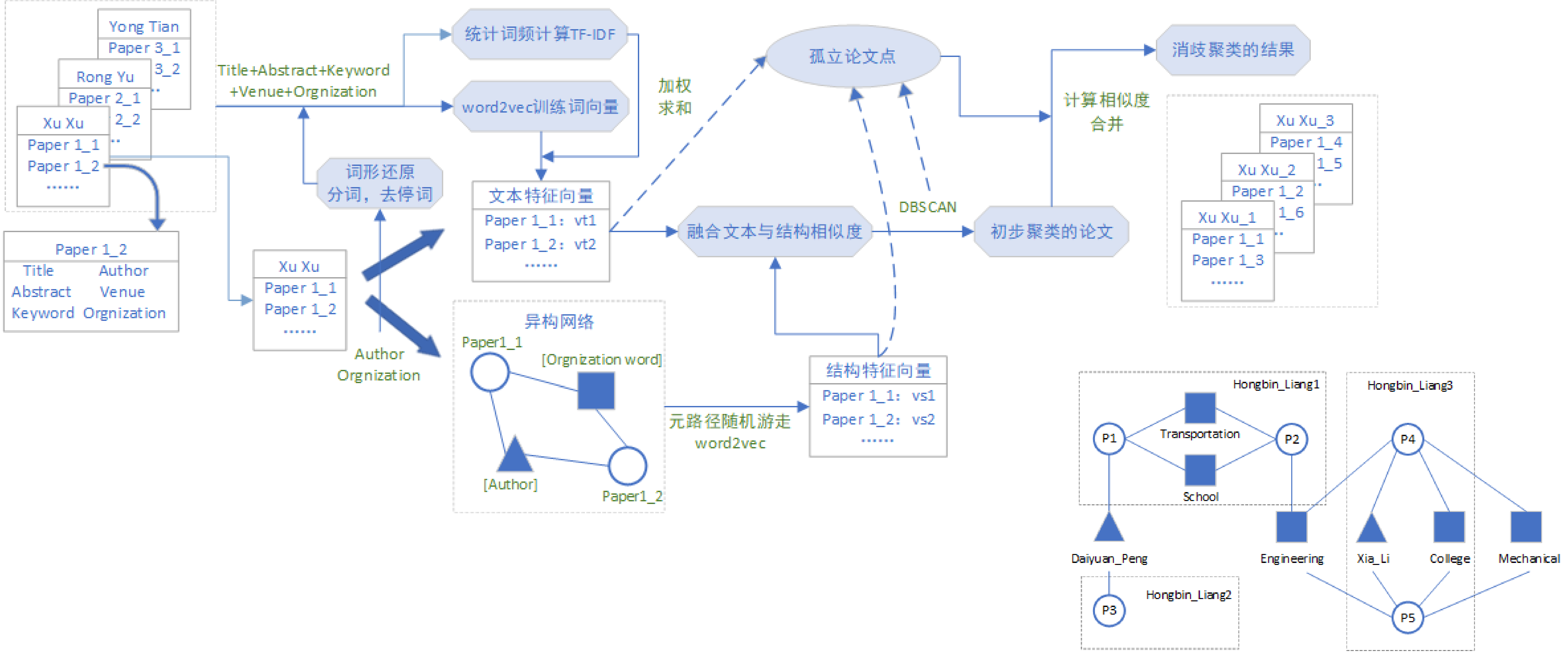
- 作者合著者、论文有共同作者2级 两个关系图
- 按边权加权随机游走长度l的路径，游走保留起点
- 用skip-gram学习表示
- HAC聚类

Author name reference	Our Method	Name Disambiguation[ND]	Graph Factorization	Deepwalk	Node2Vec	LINE	PTE
K Tanaka	**0.622**	0.571	0.334	0.450	0.304	0.398	0.173
M Jones	0.653	0.64	0.529	**0.696**	0.513	0.688	0.348
J Smith	0.498	**0.517**	0.316	0.098	0.073	0.104	0.136
Y Chen	0.515	**0.643**	0.439	0.118	0.058	0.193	0.199
J Martin	**0.782**	0.776	0.755	0.728	0.629	0.774	0.587
A Kumar	**0.615**	0.458	0.319	0.407	0.424	0.395	0.247
J Robinson	**0.698**	0.596	0.393	0.513	0.608	0.603	0.345
M Brown	**0.658**	0.602	0.478	0.481	0.211	0.633	0.269
J Lee	0.600	**0.656**	0.231	0.387	0.181	0.134	0.142
S Lee	**0.572**	0.537	0.345	0.194	0.044	0.109	0.074

Table2: Experiments results of author name disambiguation problem in CiteseerX dataset (embedding dimension = 20)

- Name Disambiguation Using Graph Node Embedding Method, IEEE, 2019

OAG比赛第一名方法



- https://www.biendata.xyz/models/category/3637,2019