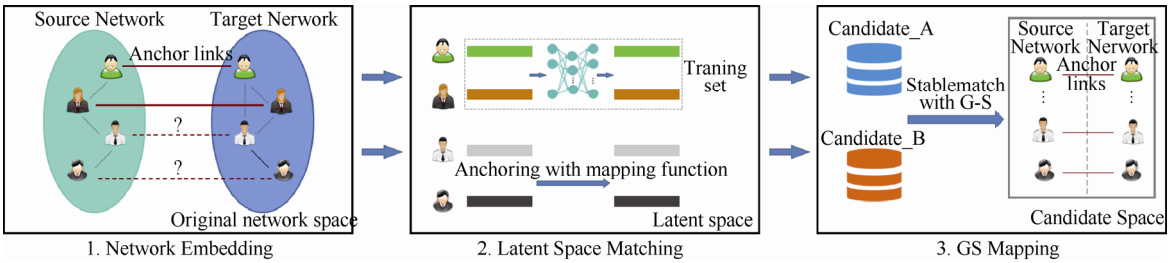


Fig.1 The schematic of IAUE model

图 1 IAUE 模型的原理图

2.1 网络 Embedding

这一阶段,通过网络的表征学习将源网络 G_s 和目标网络 G_t Embedding 成低维的向量空间.但在这一过程中,由于网络中的一些未知的“隐藏”锚链可能会造成网络的不可靠表征.基于该问题,本文将原网络和目标网络进行扩展从而确定出网络中部分“隐藏”的边:利用两个网络中可观察到的锚链信息和两个网络中的结构信息共同确定出这些“隐藏”的边.

在 IAUE 模型中,网络 Embedding 发挥了重要作用,网络的 Embedding 方法可以快速而准确地处理网络对齐问题.已有学者提出了几类著名的网络 Embedding 方法,包括 LINE[15]、Deepwalk[16]、Spectral Clustering[17].

在网络表征学习之前,先将两个网络进行扩展.通常情况下,如果两个节点在一个网络中没有连接关系,但它们在另一个网络中对应锚节点间却存在连接关系,则可以在当前网络中给这两个节点间添加一条边.

将扩展后的网络 G_s 和目标网络 G_t 分别 Embedding 成低维的向量空间.对于给定的一对顶点 v_i、v_j 和它们的低维表征 z_i、z_j,它们之间存在边的概率为

$$p(v_i,v_j)=\delta(z_i^T\cdot z_j)=\frac{1}{1+\mathrm{e}^{-z_i^T\cdot z_j}} \tag{3}$$

其中,$\delta(x)=1/(1+\exp(-x))$是一个 Sigmoid 函数.针对 $p(v_i,v_j)$的似然函数 $L(p)=\Pi p(v_i,v_j)$,有:

$$\frac{\mathrm{d}L(p)}{\mathrm{d}p}=\frac{\mathrm{d}\delta(z_i^T\cdot z_j)}{\mathrm{d}(z_i^T\cdot z_j)}=\frac{\mathrm{e}^{\left(z_i^T\cdot z_j\right)}}{\left(1+\mathrm{e}^{(z_i^T\cdot z_j)}\right)^2}>0 \tag{4}$$

所以,似然函数无解.为解决此问题并使以上概率最大化,计算条件概率 $p(v_i,v_j)$要求在整个节点集上求和,为了降低计算复杂度,采用负采样的方法[18]最大化以下目标函数:

$$\log\delta(z_i^T\cdot z_j)+\sum_{k=1}^{K}E_{v_k\propto P_n(v)}[\log(1-\delta(z_i^T\cdot z_k))] \tag{5}$$

即, $\log p(v_i,v_j)\propto\log\delta(z_i^T\cdot z_j)+\sum_{k=1}^{K}E_{v_k\propto P_n(v)}[\log(1-\delta(z_i^T\cdot z_k))]$,其中,公式(5)的第 1 项模型化已知的锚链,第 2 项表示负采样的边,并且每一个顶点被采样的概率为 $P_n(v)\sim d_v^{3/4}$ [19].另外,K 代表负采样的边数,d_v 表示顶点 v 的度.

在扩展后的源网络 G_s' 和目标网络 G_t' 中,通过最大化公式(5)以达到最小化损失函数 $O_1(1)$的值,最后采用随机梯度下降的方法学习两个网络的向量表征,为了更新节点 v_i 在网络中的向量的表征 z_i,梯度下降的计算过程为

$$\frac{\mathrm{d}O_1}{\mathrm{d}z_j}=w\frac{\mathrm{d}p(v_i,v_j)}{\mathrm{d}z_j}=\frac{\mathrm{d}\log p(v_i,v_j)}{\mathrm{d}z_j} \tag{6}$$

其中, $w=\left\|\frac{\log p(v_i,v_j)}{\log\delta(z_i^T\cdot z_j)+\sum_{k=1}^{K}E_{v_k\propto P_n(v)}[\log(1-\delta(z_i^T\cdot z_k))]}\right\|_F$,上述公式(6)的偏导数可写为

$$\frac{\mathrm{d}O_1}{\mathrm{d}z_j} = w\{1-\delta(z_i^T \cdot z_j)\}z_i - \delta(z_k^T \cdot z_j)z_k \tag{7}$$

同理,可得其他节点向量的偏导数:

$$\frac{\mathrm{d}O_1}{\mathrm{d}z_i} = w\{1-\delta(z_j^T \cdot z_i)\}z_j - \delta(z_k^T \cdot z_i)z_k \tag{8}$$

$$\frac{\mathrm{d}O_1}{\mathrm{d}z_k} = w\{1-\delta(z_k^T \cdot z_j)\}z_j \tag{9}$$

综上所述,通过公式(7)~公式(9)可学习源网络和目标网络的低维向量表征.

2.2 映射函数的学习

在这一阶段,IAUE 模型将通过监督学习得到映射函数$\Re$,监督信息为已知锚链$(v_l^s, u_n^t) \in L$的低维表征z_l^s和z_n^t,具体为,

对于$z_l^s \in Z_s$,映射函数可表示为$\Re(z_l^s;\theta)$,且该映射函数是非线性的.其中,θ是映射函数中的特征集.所以,从源网络到目标网络的损失函数为

$$\ell_m(\Re, Z_s, Z_t, L) = \min_n \left\| \Re(z_l^s;\theta) - z_n^t \right\|_F \tag{10}$$

随后,利用 BP 神经网络学习出源网络到目标网络的映射函数.利用这种方法,不仅不需要将源网络和目标网络进行线性对齐,而且也使得网络 Embedding 具有了更好的泛化能力,可以更灵活地得到网络的结构规律.

2.3 锚链识别

对于源网络中的每一个节点,能得到该节点在目标网络中的候选匹配节点集,为方便使用,该候选集中的节点一般按照与源网络中的当前节点相似度高低来排序.

在当前网络对齐的方法中,目标网络中所求出的映射节点以候选集合的形式给出.虽然一些研究结果根据节点的相似度进行了排序,但是,节点对应关系挖掘的准确度仍有待提高.基于此,本文提出了基于 G-S 算法的提升网络锚节点预测结果的方法.具体为将两个网络交换作为源网络和目标网络,最终结果得到了两个各自源网络的候选集.然后,将这两个结果作为 G-S 算法的输入进行稳定匹配,最终提升了锚链识别的准确度.IAUE 模型伪代码如算法 1 所示.

算法 1. IAUE 算法.

```
Input: Two networks A and B, a set of supervision anchor links L;
Output: candidate A, B.
01: define Mapping Function=R
02: procedure NodetoEmb(A), NodetoEmb(B)
03: Initialize A′=NodetoEmb(A), B′=NodetoEmb(B)
04: Initialize loss1(Eqs1), loss2(Eqs2), epochNum, batchNum
05: alignment(A′, B′)
06:                 for epochNo in range(1, epochNum)
07:                     loss1, loss2
08:                         for batchNo in range(1, batchNum)
09:                             Update R based on Eqs.(7,8,9)
10:                         end for
11:                 end for
12:         until convergence
13: return candidate_AMap, candidate_BMap
14: G-S(candidate_AMap, candidate_BMap)
```

```
15: end procedure
```

G-S 算法的伪代码如算法 2 所示.

算法 2. G-S 算法.

```
Input: candidate A, B;
Output: accuracy.
01: Procedure G-S(candidate_AMap, candidate_BMap):
02: if length(A)!=length(B):
03:     error 0
04: nNode=length(A)
05: candidates_AMap=readCandis(′candidates_A)
06: candidates_BMap=readCandis(′candidates_B)
07: candisMap=[candidate_AMap, candidate_BMap]
08: mappingRes=MapResultIni(nNode)
09: candidates_ARemain=range(1, len (candisMap[0])+1)
10: while len(candidates_ARemain)!=0:
11:         for candidate_A in range(1, ncandidates_A+1):
12:             if candidate_A not in candidates_ARemain:
13:                 continue
14:             if len(candisMap[0][candidate_A])>0:
15:                 candidate_B=candisMap[0][candidate_A][0]
16:             else:
17:                 refreshMapping(mappingRes, candidate_A)
18:                candidates_ARemain.remove(candidate_A)
19:                 continue
20: Output: accuracy(mappingRes[0])
21: end procedure
```

3 实验及分析

为了检验IAUE模型的性能,将IAUE模型和5种经典方法进行实验对比.本文选取的实验数据为Facebook数据集、新浪微博和豆瓣数据集.

3.1 实验方法和度量指标

为了与本文最为相关的研究工作[8]进行比较,我们同样选取了4种较为流行的做法以及Man[8]的工作作为本文算法的比较对象,在锚链识别准确率方面进行算法比较.

第1种方法为基于节点度的对齐方法:根据两个网络中节点的度进行锚链匹配,该方法属于无监督方法.

第2种方法为MNA模型[20]:利用社交网络中用户的社交信息作为识别锚链的信息,如:社会、空间、时间和文本信息等.然后利用网络中存在一部分很少的锚节点进行监督学习.

第3种方法为MAD模型[6]:通过识别网络中的一些种子节点,然后通过这些种子节点进行迭代计算,最后通过矩阵的奇异值分解来识别两个网络中的锚节点,该方法属于无监督学习方法.

第4种方法为CLF模型[7]:通过随机游走方法识别出两个网络中存在的锚节点,该方法属于监督学习方法.

第5种方法为PALE模型[8]:通过网络Embedding的方法对网络进行降维处理,然后通过多层感知器学习映射函数,最终得到源网络中的节点在目标网络中对应锚节点的候选集,属于监督学习方法.

本文的 IAUE 模型:IAUE 模型是利用 BP 神经网络学习出一个映射函数,通过映射函数得到了源网络中的锚节点在目标网络中对应的锚节点的候选集,最后通过 G-S 算法对候选集进行处理,识别出稳定的锚链结果.

IAUE 模型的最终输出结果是源网络中的锚节点在目标网络中对应的节点集,选择 F_1 和 MAP 作为实验的评价指标.

3.2 实验数据集

IAUE 模型的实验数据集为 Facebook 数据集、新浪微博和豆瓣数据集.将 Facebook 数据集、新浪微博和豆瓣数据集中的度小于 5 的节点剪枝.最终,Facebook 数据集中包含 40 710 个节点和 766 519 条边,新浪微博数据集中包含 75 387 个节点和 356 128 条边,豆瓣数据集中包含 55 387 个节点和 503 782 条边,见表 1.

Table 1 The statistics of experimental data

表 1 实验数据统计表

数据来源	Facebook	新浪微博	豆瓣
节点数	40 710	55 387	55 387
边数	766 519	656 128	803 782

3.3 Facebook 数据集的实验

在 Facebook 数据集中,由于选取两个具有关联性的异构网络较为困难,于是根据以下规则在 Facebook 数据集中选取两个子网络来模拟两个不同的社交网络:在该数据集中,为每一条边赋于一个介于[0,1]之间均匀分布的随机权值 p.如果 $p\leqslant 1-2\alpha_s+\alpha_s\alpha_c$,则将该条边舍弃;如果 $1-2\alpha_s+\alpha_s\alpha_c\leqslant p\leqslant 1-\alpha_s$,则将该条边保留在第 1 个子网络中;如果 $1-\alpha_s\leqslant p\leqslant 1-\alpha_s\alpha_c$,则将该条边保留在另一个子网络中;否则,对于其他情况,该条边会被同时保留在两个网络中.其中,α_s 是从原始网络采集边的速率,反映了网络的稀疏水平,α_c 代表网络重叠度(即会有一些节点同时存在于两个网络中).基于这样的采样策略,得到两个子网络.在实验过程中,选取其中一个子网络作为源网络,记为 G_s,另一个子网络作为目标网络,记为 G_t,在两个网络中,利用已知的锚链信息进行监督学习,然后通过 G-S 算法识别未知的锚链,已知的锚链信息记为α_l.通过多次实验,其结果表明:IAUE 模型在设定不同的α_s 和α_c 时,它的性能都优于其他方法.

为了评价本文提出的 IAUE 模型,设定了两种不同的情形,并将它与 5 种网络对齐算法进行实验对比.同时,设定α_l=3%,即训练集中含有 3%的锚链信息.首先,设定α_c=0.9,对不同的α_s=[0.5,0.6,...,0.9]进行实验,实验结果见表 2.然后,设定α_s=0.6,对不同的α_c=[0.5,0.6,...,0.9]进行实验,实验结果见表 3.

Table 2 Experimental results of identifying anchor links under different α_s=[0.5,0.6,...,0.9], α_c=0.9

表 2 α_c=0.9,不同α_s=[0.5,0.6,...,0.9]的锚链识别方法的实验结果

评价指标	方法	网络稀疏度α_s				
		50%	60%	70%	80%	90%
F_1	Degree	0.092 2	0.093 2	0.094 5	0.094 7	0.095 4
	MAD	0.389 9	0.389 0	0.389 3	0.390 4	0.392 2
	MNA	0.4262±0.0011	0.4290±0.0016	0.4370±0.0015	0.4365±0.0011	0.4390±0.0012
	CLF	0.8693±0.0006	0.8724±0.0012	0.8734±0.0013	0.8786±0.0009	0.8820±0.0022
	PALE	0.8936±0.0005	0.8943±0.0012	0.8966±0.0011	0.9009±0.0012	0.9012±0.0011
	本文 IAUE	**0.9128±0.0003***	**0.9147±0.0008***	**0.9173±0.0010***	**0.9196±0.0011***	**0.9205±0.0010***
MAP	Degree	0.092 8	0.097 9	0.098 3	0.098 7	0.099 1
	MAD	0.393 1	0.403 2	0.409 7	0.411 2	0.412 8
	MNA	0.4571±0.0002	0.4573±0.0015	0.4583±0.0013	0.4591±0.008	0.4594±0.0007
	CLF	0.8834±0.0005	0.8835±0.0008	0.8898±0.0004	0.8912±0.0011	0.8915±0.0004
	PALE	0.9100±0.0008	0.9207±0.0009	0.9224±0.0011	0.9228±0.0005	0.9237±0.0008
	本文 IAUE	**0.9226±0.0006***	**0.9287±0.0005***	**0.9301±0.0010***	**0.9310±0.0008***	**0.9346±0.0008***

从表 2 中的数据可以看出,使用节点度去识别锚链的方法性能最差,它的 F_1 指数和 MAP 指数都是最低的.当α_c=0.9 时,MAD 的 F_1 值和 MAP 值分别达到了大约 0.39 和 0.41,这是因为 MAD 方法没有使用已知的锚链作监督进而影响了结果的准确性.对于 MNA、CLF、PALE 和本文的 IAUE 方法,它们都属于监督学习的方法,为了评价这几种方法的性能并降低实验结果的相对误差,进行了 10 次实验并取平均值作为最终实验结果,且设定

训练集中的锚链数量α_l=3%.从表 2、表 3 可以看出,监督方法的结果明显优于非监督方法,即 MNA、CLF、PALE 和 IAUE 这 4 种方法的性能明显优于 MAD 方法,并且在这 4 种监督学习方法中,IAUE 模型的 F_1 值和 MAP 值都是最高的.

Table 3 Experimental results of identifying anchor links under different α_c=[0.5,0.6,...,0.9], α_s=0.6

表 3 α_s=0.6,不同α_c=[0.5,0.6,...,0.9]的锚链识别方法实验结果

评价指标	方法	网络重叠度α_c				
		50%	60%	70%	80%	90%
F_1	Degree	0.008 9	0.012 8	0.033 4	0.066 4	0.093 2
	MAD	0.102 0	0.152 3	0.202 1	0.333 7	0.389 0
	MNA	0.1340±0.0012	0.1888±0.0010	0.2170±0.0013	0.3521±0.0015	0.4290±0.0016
	CLF	0.2940±0.0010	0.3823±0.0012	0.5510±0.0012	0.7350±0.0008	0.8724±0.0012
	PALE	0.3789±0.0005	0.4518±0.0009	0.5914±0.0010	0.7714±0.0011	0.8943±0.0012
	本文 IAUE	**0.4250±0.0003***	**0.5336±0.0006***	**0.6697±0.0006***	**0.7935±0.0010***	**0.9123±0.0010***
MAP	Degree	0.010 2	0.013 3	0.035 8	0.070 4	0.097 9
	MAD	0.132 1	0.163 3	0.231 2	0.344 9	0.403 2
	MNA	0.1450±0.0012	0.2498±0.0010	0.2923±0.0011	0.3978±0.0012	0.4573±0.0015
	CLF	0.3245±0.0007	0.4133±0.0009	0.5998±0.0014	0.7732±0.0010	0.8835±0.0008
	PALE	0.4197±0.0010	0.4845±0.0007	0.6224±0.0012	0.8118±0.0010	0.9207±0.0009
	本文 IAUE	**0.4526±0.0006***	**0.5342±0.0008***	**0.6597±0.0010***	**0.8562±0.0008***	**0.9432±0.0010***

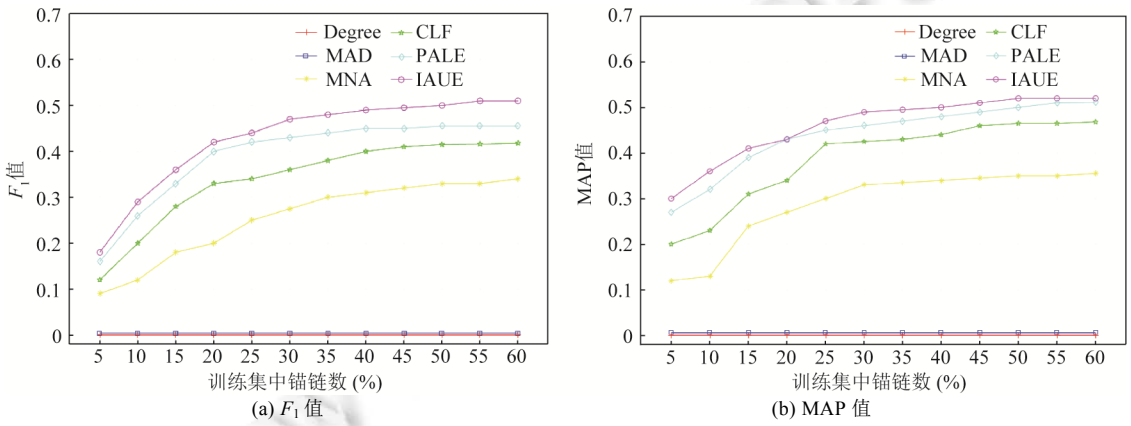
MNA 方法仅考虑了网络中节点的相似度,它表现出的性能比 CLF 方法和 IAUE 方法都低.同样,它的结果也再次说明了网络的结构信息对于网络对齐问题的重要性,并且本文所使用的网络 Embedding 方法是一种有效的能够较为准确地得到网络结构规律的一种途径和方法.

在本文中,通过实验验证了α_l 从 0.5%~5%的过程中对实验结果造成的影响.从最终实验结果可以看出,当 0.5%≤α_l≤1.5%时,IAUE 模型的性能表现出很快的增长趋势.相比于 PALE 模型,本文的 IAUE 模型可以使用更少的锚链信息学习,但却可以达到更高的准确率.并且,随着两个网络中节点重叠度α_c 的提高,实验结果会越来越好,这说明了α_c 的值对于锚链识别的准确性也是一个重要的指标.

3.4 新浪微博-豆瓣的实验

IAUE 模型的第 2 个实验数据集为新浪微博和豆瓣网数据集.新浪微博和豆瓣属于两个不同的社交网络,将这两个网络中节点度小于 5 的节点进行剪枝,数据集中包含 875 对锚节点,这些锚节点将作为模型学习的监督信息.

将训练集中的锚链数量α_l 设定为从 5%~60%,将其余的锚链作为测试数据.对于每一个α_l,进行 10 次实验并取平均值作为实验结果.实验结果如图 2(a)和图 2(b)所示,由实验结果可以看出,IAUE 模型在锚链识别问题上的性能均优于其他几种方法.



(a) F_1 值 (b) MAP 值

Fig.2 The comparison of experimental performance

图 2 实验性能对比

4 结 语

本文提出了一种有效的网络对齐模型:IAUE 模型.该模型的方法属于监督学习的方法,相比无监督的方法,IAUE 模型利用网络表征学习,并结合 BP 神经网络和 G-S 算法,极大地提高了锚链识别的准确性.在社交网络 Facebook 数据集、新浪微博和豆瓣数据集上,IAUE 模型与 MNA、MAD、CLF、PALE 以及基于度的方法进行了实验对比.实验结果表明,本文提出的方法识别锚链效果优于其他 5 种方法,本文的后续工作将包括增加其他属性特征,并将这些特征更科学地加以结合,等等.

References:

[1] Backstrom L, Dwork C, Kleinberg J. Wherefore art thou r3579x?: Anonymized social networks, hidden patterns, and structural steganography. In: Proc. of the Int'l Conf. on World Wide Web. ACM, 2007. 181–190.

[2] Dong Y, Tang J, Wu S, *et al*. Link prediction and recommendation across heterogeneous social networks. In: Proc. of the Int'l Conf. on Data Mining. IEEE, 2013. 181–190.

[3] Zhang J, Yu PS, Zhou ZH. Meta-Path based multi-network collective link prediction. In: Proc. of the 20th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. New York: ACM, 2014. 1286–1295. [doi: 10.1145/2623330.2623645]

[4] Hu L, Cao J, Xu G, *et al*. Personalized recommendation via cross-domain triadic factorization. In: Proc. of the Int'l l World Wide Web Conf. Steering Committee. 2013. 595–606.

[5] Novak J, Raghavan P, Tomkins A. Anti-Aliasing on the Web. In: Proc. of the Int'l Conf. on World Wide Web. ACM, 2004. 30–39.

[6] Li CY, Lin SD. Matching Users and Items Across Domains to Improve the Recommendation Quality. ACM, 2014. 801–810.

[7] Zhang J, Yu PS. Integrated anchor and social link predictions across social networks. In: Proc. of the Int'l Conf. on Artificial Intelligence. AAAI Press, 2015.

[8] Man T, Shen HW, Jin XL, Cheng XQ. Predict anchor links across social networks via an embedding approach. In: Proc. of the 25th Int'l Joint Conf. on Artificial Intelligence (IJCAI-16). 2016.

[9] Zhang Y, Tang J, Yang Z, *et al*. COSNET: Connecting heterogeneous social networks with local and global consistency. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2015. 1485–1494.

[10] Kollias G, Mohammadi S, Grama A. Network similarity decomposition (NSD): A fast and scalable approach to network alignment. IEEE Trans. on Knowledge & Data Engineering, 2012,24(12):2232–2243.

[11] Bayati M, Gerritsen M, Gleich DF, *et al*. Algorithms for large, sparse network alignment problems. In: Proc. of the 9th IEEE Int'l Conf. on Data Mining (ICDM 2009). Miami, 2009. 705–710.

[12] Gale D, Shapley LS. College admissions and the stability of marriage. American Mathematical Monthly, 2013,120(69):9–15.

[13] Grover A, Leskovec J. node2vec: Scalable feature learning for networks. In: KDD. 2016. 855–864. [doi: 10.1145/2939672.2939754]

[14] Wang D, Cui P, Zhu W. Structural deep network embedding. In: Proc. of the ACM SIGKDD Int'l Conf. ACM, 2016. 1225–1234.

[15] Tang J, Qu M, Wang M, *et al*. LINE: Large-Scale information network embedding. In: Proc. of the Int'l World Wide Web Conf. Steering Committee. 2015. 1067–1077.

[16] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online learning of social representations. In: Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. ACM, 2014. 701–710.

[17] Tang L, Liu H. Leveraging social media networks for classification. Data Mining and Knowledge Discovery, 2011,23(3):447–478.

[18] Mikolov T, Le QV, Sutskever I. Exploiting similarities among languages for machine translation. Computer Science, 2013.

[19] Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 2013,26:3111–3119.

[20] Kong X, Zhang J, Yu PS. Inferring anchor links across multiple heterogeneous social networks. In: Proc. of the ACM Int'l Conf. on Information & Knowledge Management. ACM, 2013. 179–188.

王宁(1992－),男,山西大同人,硕士,主要研究领域为大数据,社会网络计算.

王莉(1971－),女,博士,教授,博士生导师,CCF 高级会员,主要研究领域为人工智能,社会网络计算.