图对齐课题研究报告*

杨博源[1]，蔡鸿伟[2]，杨永祥[3]

[1](信息科学技术学院　2000012909)

[2](城市与环境学院 1800013236)

[3](信息科学技术学院 2000010801)

摘　要：随着网络的普及，各类网络社交平台逐渐步入人们的视野，人们为了满足不同的社交活动需求往往会使用多个社交软件，因此跨社交网络环境下的用户匹配识别问题逐渐成为了热门话题。基于用户的账号关联信息以及在社交平台上的活动信息，我们在利用自然语言处理中的 word2vec 和 LDA 模型将用户的文本数据进行了向量嵌入，将其作为对结点相似度计算的另一依据；此外我们还改进了在图上扩展游走的策略，改进后的算法相比现有的逐步扩展子图的迭代时间代价更低；最后我们在真实的数据集上集中验证了算法的有效性。

关键词：社交网络；图对齐；MAUIL；REGAL；相似度计算

Figure alignment Topic Research Report

YANG Bo-Yuan[1], CAI Hong-Wei[2], YANG Yong-Xiang[3]

[1](School of Electronics Engineering and Computer Science, 2000012909)

[2](College of Urban and Environmental Sciences, 1800013236)

[3](School of Electronics Engineering and Computer Science, 2000010801)

Abstract: With the popularity of the Internet, various social networking platforms have gradually come into people's vision. In order to meet the needs of different social activities, people often use multiple social software. Therefore, the problem of user matching and identification in the cross social networking environment has gradually become a hot topic. Based on the user's account association information and the activity information on the social platform, we use the word2vec and LDA models in natural-language-processing to embed the user's text data as another basis for calculating the node similarity; In addition, we also improve the strategy of extending the walk on the graph. Compared with the existing gradually expanding subgraphs, the improved algorithm has lower iteration time cost; Finally, we verify the effectiveness of the algorithm on real data sets.

Key words: social networks; figure alignment; MAUIL; REGAL; similarity calculation

随着网络的普及，各类网络社交平台逐渐步入人们的视野，人们为了满足不同的社交活动需求往往会使用多个社交软件，例如豆瓣为用户提供了图书、电影、音乐交流分享服务，微博为用户分享日常活动包括分享动态提供服务，知乎为用户提供了问答形式的互动服务，同一用户也因此会在不同的网站注册账号，因此跨社交网络环境下的用户匹配识别问题逐渐成为了热门话题。一般而言，我们讨论的简单抽象模型问题即带信息的结点图上的对齐问题。

针对网络对齐问题,在早期研究工作中,研究者们主要利用用户 E-mail、用户真实姓名等信息进行识别,虽然依据 E-mail 和真实姓名能够精确匹配用户,但在真实社交网络中,E-mail 和真实姓名由于隐私保护的原因,通常难以获取.因此,现阶段工作主要集中于利用：(1) 用户属性信息,如用户头像、用户喜好等；(2) 用户行为信息,如发帖关键字、用户行为轨迹等；(3) 用户结构信息,如用户朋友关系、用户与帖子的评论关系等。虽然现有方法具有良好的准确性,但真实社交网络通常面临用户数据匿名化严重、部分用户数据难以获取等问题,且现有公开数据也面临数据缺失、数据不一致、数据分布不均、数据异构等问题。

为简化讨论，本文将在结点有信息的图上进行对齐问题的讨论，即给定无向图 G=<V, E>，其中结点表示实际社交网络中的账号，每个账号有信息 Li，表示该节点在社交网络中的账号信息活动，E 表示两个账号有

互相关注或者互动历史。现有的图对齐算法对文字嵌入的研究较少，主要是基于机器学习的神经网络和基于图数学结构本身对结点估值的两大类方法。前者的优点是正确率较高，后者的优点是时间效率表现优秀，我们的工作主要是：(1)将文本信息转化为向量嵌入到相似度计算；(2)改进了图的迭代规则，进一步提升算法效率；(3)在实际的数据集上对算法进行了一定的分析。

算法设计相关工作

首先我们对当前图对齐算法的两个主要方向——基于机器学习的 GNN 模型和基于图结构的矩阵变换进行了文献阅读总结，选取了 iMap: Incremental Node Mapping between Large Graphs Using GNN[1]和 REGAL: Representation Learning-based Graph Alignment[2]两篇论文进行分析讨论，得出以下几点启发：

1、从 iMAP 最终得到模型的实验结果来看，可能被对齐的节点是与锚点紧密相关的，因此只提取靠近锚的节点进行扩展迭代，因此在设计算法时可以利用这种局部性，从既定锚点出发扩展有助于得到相似度较高的锚点对。

2、随机游走策略可能会导致很多不必要的操作，且该策略主要目的是为了构建适合 GNN 训练的输入，因此在算法不基于机器学习时可以替换为指定方向的图扩展。

3、基于纯数学理论的图论推导能够在时间复杂度上得到提升，但所需的空间较大，且多项式时间算法的次数可能很高，在较小的子图上表现优异。

其次，我们给出了在未进行文字嵌入时的算法运行过程，即根据图结构本身的算法，此时相似度的计算方法采用 REGAL 论文给出的矩阵上的相似度计算方法。算法描述如下。

数据结构：关于锚点对的集合 S，最终的输出即为图对齐结果。

所用符号：p 表示采用 p 跳展开，k 表示每次迭代选取的锚点对数，b 表示每次检验的标准值，Sim(u, v)表示点 u, v 的相似度。

算法过程：

Step1: 从初始给定的锚点对出发，在两图中分别以初始锚点为中心外跳 p 层，即得到当前聚焦的子图 G1，G1'。在 G1 与 G1'上调用 REGAL 算法得到相似度最高的 k 对锚点，将它们置于集合 S。

Step2: 选取集合 S 中相似度最高的一对锚点作为这一步迭代的锚点，在两图中分别以它们为中心外跳 p 层，得到当前聚焦的子图 G2，G2'。

Step3：在 G2，G2'上首先对当前 S 中的点对(如果在对应的子图中)计算相似度，若相似度小于 b 则将其从 S 中删除。随后调用 REGAL 算法得到相似度最高的 k 对锚点，重复 Step2-Step3，直至每个点都被聚焦过的子图覆盖至少一次，输出 S。

最后，经过简单的预实验分析，我们选择 p 值为 4，k 值为 5~8，b 值为 0.9 作为该算法的具体数值，这将会应用到我们后续的模型中。

文本嵌入与相似度计算的相关工作

首先，我们将节点自身的属性信息分为三类，单词级，短语级，段落级，其中单词级一般是节点的标识或者用户名，短语级往往是位置、从属、身份、职业认证、教育经历等信息，段落级则是消息内容，文章摘要等大段文字。对于输入的数据，我们按照上述指标将其分为这三类语料库，然后进行恢复原型，去除多余符号助词感叹词，分词等操作，然后借助 nltk 和中文维基百科的语料库各自预训练中英文模型以供后续使用。单词级嵌入使用了 gensim 库的 model 模型，短语级嵌入使用了流行的 Word2Vec 模型[3]，段落级嵌入则使用 LDA 主题模型。

除了节点属性外，不同节点之间的关系同样蕴含着重要信息，传统的图嵌入算法一般基于近邻相似假设，即两个节点的共同邻居越多则越相似。然而在图对齐场景中，两个不近邻的顶点也可能拥有很高的相似性，可能这些节点在邻域中的角色相似，是不同社交网络中的同一个用户。Struc2vec 就是针对捕捉节点的结构化角色相似度提出的模型[4]，我们使用此模型进行图结构的嵌入。经过上述过程，我们输入的带有属性和结构

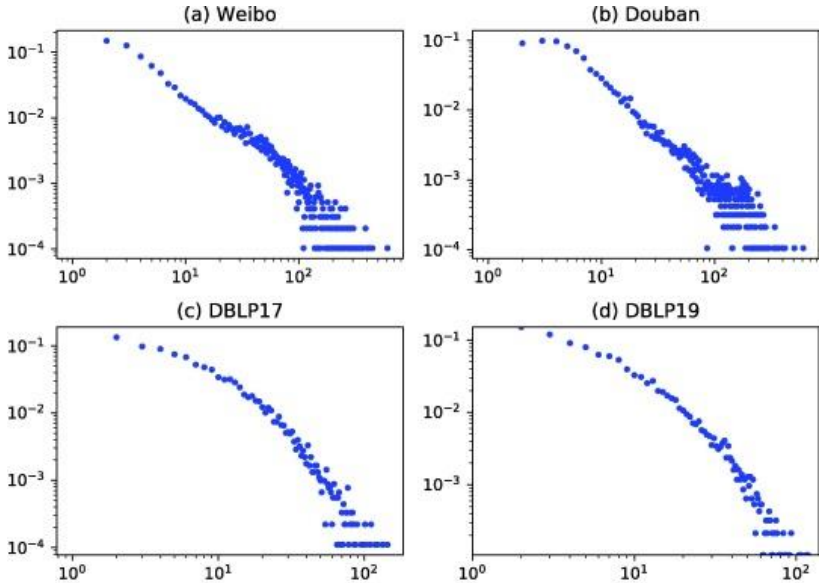
信息的两个不同社交网络原始数据集变成了两个向量集，即每个节点变成一个长为 400 维的向量，单词级、短语级、段落级和图结构的嵌入各占 100 维。

得到了每个节点的向量数据后，接下来我们需要确定两个节点的相似度函数。根据此前的研究，认为两个向量集中对应节点之间为线性关系是合理的[5]，而正则化典型相关分析（RCCA）在最大化线性相关关系的分析中表现出色[6]，因此我们使用了 RCCA 模型计算不同节点的相似度。

最后，我们简述一下训练过程。第一步，我们将已知对应关系的部分锚点对用于训练 RCCA 模型，来找到两个向量集的共同空间投影；第二步，我们将剩下的锚点分别用此模型计算其投影并计算距离它最近的另一个向量集的点，检测是否为对应点的成功率，以验证 RCCA 模型的合理性。第三步，我们将全部锚点用于训练 RCCA 模型，然后在两个向量集各自任给一点，就可以通过计算它们的空间投影之间的距离来表示其相似度。

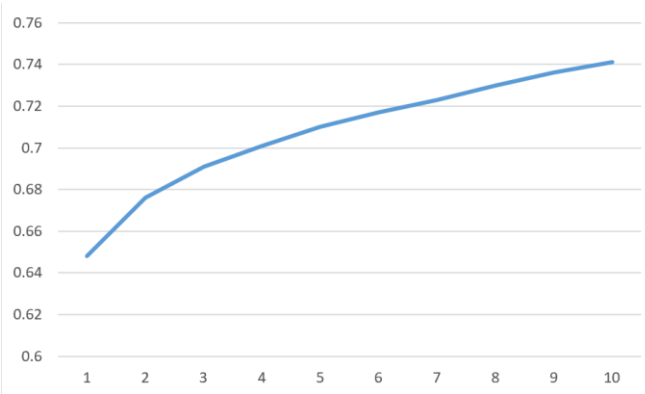
实验数据与分析

我们从网络是搜集到了两种数据集，一是来自微博和豆瓣两个社交平台的用户转发消息数据，二是 DBLP 计算机科学书目网络不同时期的网络备份，这些数据的基本信息如下。



Datasets	Networks	#Users	#Relations	Min. degree	Ave. degree	Max. degree	Ave. coeff.	#Matched pairs
Social	Weibo	9,714	117,218	2	12.1	607	0.112	1,397
networks	Douban	9,526	120,245	2	12.6	608	0.101	
coauthor	DBLP17	9,086	51,700	2	5.7	144	0.280	2,832
networks	DBLP19	9,325	47,775	2	5.1	138	0.322	

以 DBLP 数据集为例，我们以 500 组锚点对训练 RCCA 模型，然后取 200 组锚点计算所得 k 个最相似的点中正确锚点的位置并以此打分，结果如下图所示。



随着 k 的增加，找到正确锚点的概率和整体评分也随之增大，从 0.65 上升到 0.74，这说明我们的 RCC 模型相对可靠，能够大概率找出正确的锚点对，因此也可用于探索未知的点之间的关系。

对于我们给出的子图迭代算法，以下给出了一个理论上的时间复杂度分析。

设图中顶点度数的平均值为 d，每次聚焦的子图点与点之间的距离不大于 p，因此子图平均至多有 d^p 个顶点，算法 REGAL 运行最坏情况下复杂度为 O(N)，此处N=d^P，算法的其他操作时间阶数显然小于此。设总顶点数数为 M，则每次迭代的聚焦至少偏离中心锚点 1，因此总体时间复杂度为 O($M/d \cdot d$^p) = O($M \cdot d$^(p−1))。一般而言 p 的取值不大于 4，因此该时间复杂度至多为 O(M^4)。

探究缺陷和未来的工作

第一，在我们的 RCCA 模型之中，只考虑了使用已知的锚点对进行训练，而没有考虑负样本[7]。这种情况下，两个相似但是不对应的节点就会影响模型的性能，RCCA 模型无法区分真正对应节点导致的高相似度和其他因素造成的高相似度的区别，那么最终根据相似度来对齐图中节点的办法正确率自然会受到影响。

第二，我们的实验分析中缺少绝对相同数据集下不同算法的绝对运行时间统计分析，仅从理论上得到多项式的时间不能有效体现这种质变。

因此，在之后的模型建立和分析之中，我们需要充分考虑负采样的作用和质量，使其能够更好地优化模型性能，节约计算量，加快运行速度；在实验部分，我们需要更多的实验样本以及时间计算相关代码以辅助完成我们的算法测评。

References:

[1] Xia Y, Gao J, Cui B. iMap: Incremental Node Mapping between Large Graphs Using GNN. 2021.

[2] Heimann M, Shen H, Safavi T, et al. REGAL: Representation Learning-based Graph Alignment[C]// the 27th ACM International Conference. ACM, 2018.

[3] Mikolov T, Corrado G, Kai C, et al. Efficient Estimation of Word Representations in Vector Space[C]// Proceedings of the International Conference on Learning Representations (ICLR 2013). 2013.

[4] Ribeiro L, Savarese P, Figueiredo D R. struc2vec: Learning Node Representations from Structural Identity[C]// arXiv. arXiv, 2017.

[5] Li C, Wang S, Yu P S , et al. Distribution Distance Minimization for Unsupervised User Identity Linkage[C]// Conference on Information and Knowledge Management. ACM, 2018.

[6] D Hardoon, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods[J]. Neural computation, 2004, 16(12):p. 2639-2664.

[7] Wang X, Xu Y, He X, et al. Reinforced Negative Sampling over Knowledge Graph for Recommendation[C]// WWW '20: The Web Conference 2020. 2020.