一种改进的针对合著关系网络的链接预测方法*)

郭景峰　王春燕　邹晓红　赵鹏飞　张　健
(燕山大学信息科学与工程学院　秦皇岛 066004)

摘　要　主要针对那些实体类标号属性未知的社会网络进行链接预测。由于实体的类标号属性与具体的社会网络有关,因此具体解决对作者之间合著关系网络图的链接预测问题。首先,给出了合著关系图的结构表示,然后把一个作者是否是多产的定义为合著关系图中作者实体的类标号属性。另外,还提出了一种改进的利用有指导学习进行链接预测的方法。在改进的链接预测方法中为每对作者新引入了一个特征属性——是否至少有一个是多产的。当所要预测的合著关系图中作者实体的类标号属性不完全已知时,用改进后的 ICCLP 算法对合著关系进行预测,以提高链接预测的性能。改进后的 ICCLP 算法中采用上面提到的改进后的链接预测方法。

关键词　链接预测,类标号属性,ICCLP,合著,多产的

Improved Link Prediction Method for Co-authorship Network

GUO Jing-feng　WANG Chun-yan　ZOU Xiao-hong　ZHAO Peng-fei　ZHANG Jian
(College of Information Science and Engineering, Yanshan University, Qinhuangdao 066004, China)

Abstract　This paper specially predicted links in the social network where the class labels of the objects are unknown. Since that the information of the objects' class labels is related to concrete social network, it specially solves the problem of link prediction for co-authorship network. At first, the structure of the co-authorship network was formulated. Then the class labels of the author objects in this network were defined as either prolific or not. And one improved supervised learning method for predict co-authorship links was proposed. In our supervised learning method, one feature (either or both prolific) that refers to the class labels of the author objects was increased. When the class labels of the author objects are unknown, the improved ICCLP which uses the improved supervised learning for link prediction is used. The experiment results show that using improved ICCLP for predicting co-authorship links in incomplete co-authorship network can get better performance.

Keywords　Link prediction, Class label, ICCLP, Co-authorship, Prolific

1　引言

近几年来,社会网络分析得到越来越多的关注。从数据挖掘的角度看,社会网络分析又称链接挖掘[1]。而链接预测是该领域非常重要的一个研究方向。链接预测就是识别两个对象之间是否存在联系。链接预测问题包括两方面的含义:一方面可以理解为识别实际存在但当前网络中并不可见的链接;另一方面可理解为基于时刻 t 的社会网络状态预测在时刻 $t+1$ 将会在网络中增加哪些链接。这两种情况下,链接预测问题均被转化成分类问题来解决。链接预测是比较困难的,因为对象之间特定链接的先验概率通常非常低[2]。本文所要做的工作就是改进现有的链接预测方法,以提高链接预测的性能。

通常,实体分类和链接预测都是被独立研究的。并且一般情况下,实体分类是在实体之间的链接信息完全已知的情况下进行的,同样链接预测也是在实体的属性完全已知的情况下进行的。然而,现实世界的社会网络,实体的属性和链接信息并不都是完全已知的。特别是当实体的类标号属性和链接是否存在均未知时,需要利用 Bilgic 和 Mark Namata 提出的 ICCLP 算法[3],即通过迭代的交叉进行集体分类和链接预测,同时实现对实体分类和预测链接的存在性。文献[3]中的实验表明对于不同类型的社会网络,利用 ICCLP 算法都比单一地进行实体分类和链接预测性能更好。然而文献[3]中提到在 ICCLP 中可以用任何的集体分类算法和链接预测算法。这种说法不太恰当。当然,所有的集体分类算法都会用到实体之间的链接信息,但并不是所有的链接预测算法都会用到相关实体的类标号属性信息。已有的大部分链接预测算法[4-6]并没有用到实体的类标号属性。当把没有用到相关实体类标号属性的链接预测算法用到 ICCLP 中时,集体分类的结果并不会给链接预测提供帮助,ICCLP 算法也不会像预期的那样迭代进行。正像文献[3]中的实验结果提到的那样,ICCLP 的链接预测部分比单一地进行链接预测,性能提高不明显。因此,可以通过改进已有的链接预测算法使其用到相关实体的类标号属性,再将这样的链接预测算法用到 ICCLP 中,以提高链接预测的性能。

由于想把相关实体的类标号属性用到链接预测算法中,故应该考虑某一具体的社会网络,本文考虑从科学刊物数据集中得到的作者合著关系网络。尽管以前有人用过这个社会

*)基金项目:国家自然科学基金项目(60673136)。**郭景峰**　博士,教授,硕士生导师,CCF 会员,主要研究领域为数据库理论及应用、数据挖掘技术;**王春燕**　硕士研究生,研究方向为链接预测。**邹晓红**　硕士,副教授,研究方向为多关系数据挖掘;**赵鹏飞**　硕士研究生,研究方向为数据库理论及应用。

网络做实验，但并没有给出这个社会网络图的具体结构。本文将确切表示合著关系图的结构，并且将改进一种利用有指导学习进行链接预测的方法。

2 相关工作

Bilgic 和 Mark Namata 提出的 ICCLP 算法，通过迭代进行集体分类和链接预测同时实现对实体分类和链接预测。文献[3]中的实验表明，ICCLP 的实体分类部分相对于单一的实体分类性能显著提高，而链接预测部分的性能却不比单一进行链接预测提高多少。如果把用到实体类标号属性的链接预测算法用到 ICCLP 中，链接预测的性能将会提高。

但是大多数链接预测算法并没有用到相关实体的类标号属性。Pavlov 和 Pavlov[7] 从过去的合作关系图中提取出一些结构属性构成训练数据集，然后用指导学习算法在训练数据集上训练得到分类模型，即链接预测模型，他们只用到了基于图的结构属性。Madadhain 和 Smyth[8] 提到用来进行链接预测的特征属性包括实体属性和基于图的结构属性。Hasan 等人[9] 提出了为进行链接预测提供帮助的一系列特征属性集，他们提到的特征既涉及到实体属性，又涉及到基于图的结构属性。但是所有这些算法都没有用到相关实体的类标号属性。本文想把相关实体的类标号属性作为通过分类解决链接预测的一个特征属性。具体详细实现如下。

3 在利用有指导学习实现链接预测的方法中引入实体类标号属性

社会网络可以用图表示，节点对应实体，边则对应表示实体间关系的链接。在用图描述的社会网络中，可以有不同类型的实体，实体也可以有自身的属性及类标号属性。同样，社会网络图中也可以有不同类型的链接，链接也可以有自身的属性，并且链接可以是有向的或无向的。就本文要分析的合著关系网络图而言，有两种类型的实体：文章实体和作者实体。为了下面分析的需要，文章实体有其自身的一些属性，诸如题目、关键词、作者、出版年份及所属学科。作者实体有其类标号属性——是否多产。这个图中有两种类型的链接。若两个作者实体一起写过一篇或多篇文章，那么构成对应两个节点之间的一种合著关系链接。作者实体节点和他所写的文章实体节点之间构成另外一种链接——著有，这种链接关系是有向的。合著关系网络图的主要结构如图 1 所示。

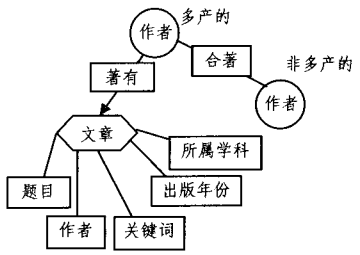

图 1 合著关系网络图的主要结构例示

3.1 为有指导的学习准备训练数据集

为了进行链接预测，把文章的出版年份划分成不重叠的两个范围，一部分作为训练年份，另一部分作为测试年份。然后从合著关系网络图中提取要得到的训练数据集。选取在训练年份中没有合著过的一对作者作为训练样本，根据这对作者在测试年份中是否会合著，将对应训练样本的类标号属性设为真或假。由此将链接预测问题转化成了二元分类问题。这个分类问题可以通过对由有效的特征属性组成的训练数据集进行有指导的学习来解决。

因为已经有了很多可以借鉴的有指导学习分类算法，关键问题在于构造用于分类算法训练的训练数据集。上面已经给出了训练样本的具体定义，因此关键在于为训练样本选取合适的特征属性。因为最终目的是链接预测，所以所选取的特征属性应当能描述作为训练样本的两个节点之间的接近程度。

可以采用 Madadhain 等人[8] 和 Hasan 等人[9] 提出的一些特征属性，诸如关键词匹配数量、邻居数、最短距离、共同邻居数目以及 Jaccard's coefficient，$Katz_\beta$，本文另外增加了一个特征属性。首先，把是否为多产的作为作者实体的类标号属性。如果一个作者是多产的，那么他很可能从事跨学科研究，继而在整个合著关系网络图中他也可能会建立比较丰富的链接关系。由此可以推得对应训练样本的一对作者中一个或是两个都是多产的，那么这对作者在测试年份合著的可能性要比随机的一对作者可能性更大。这是基于这样一个事实：随着图的演变，图中度数越大的节点其度数增加得更多。于是本文把是否至少有一个是多产的作为另外增加的特征属性。如果训练样本对应的一对作者有一个或两个都是多产的，那么该特征属性取值为真，相应的训练样本类标号属性取值为真的可能性增加，即对应的一对作者合著的可能性增加。为了计算该特征属性的值，作者实体的类标号属性必须是已知的。提到特征属性的取值，还必须专门强调一下“邻居数”的取值。这个特征与单个作者节点有关，而特征属性是对应于一对作者节点的。由于一个有广泛合著链接关系的人更有可能与其他人产生新的合著链接，因此这个特征属性的取值为两个作者节点的邻居数目的较大值。至于其他特征属性的取值，可以从文献[8,9]中得到。需要强调的是，这里涉及到的邻居关系以及节点之间的路径信息都对应于作者节点间的合著链接关系。而图中的文章节点及其属性都是用来为计算训练样本的特征属性值提供帮助的。由上可知，对于满足训练样本条件的每一对作者节点，都可以计算它的上述特征属性值。对于每一个训练样本，其特征属性值的集合构成了一个特征向量。训练样本根据其对应的一对作者是否在测试年份合著其类标号属性取真或假。对应于一个训练样本的特征向量及其类标号属性值构成训练数据集的一个训练元组。到目前为止，已经得到了用作分类的训练数据集。

3.2 应用已有的分类算法构造链接预测模型

在 3.1 节中，已经将链接预测问题转化成了分类问题，并且也得到了用于分类的训练数据集，现在只需要用已有的一些分类算法分析训练数据集来构造分类器，即链接预测模型。Madadhain 等人[8] 和 Hasan 等人[9] 用了一些分类算法来构造分类器，诸如支持向量机(SVMs)[10]、决策树、k 最近邻分类法(K-Nearest Neighbors)及朴素贝叶斯等。Hasan 等人[9] 用实验表明大多数的分类模型都可以用来解决链接预测问题，但就综合性能考虑，SVM 性能最好。因此，在本文的实验中，用线性 SVM 分析 3.1 节得到的训练数据集，从而得到分类器，即链接预测模型。

4 将上述与实体类标号属性有关的链接预测算法应用到 ICCLP 中

第 3 节中的链接预测算法是在文章实体的所有属性及作者实体的类标号属性都已知的前提下提出的。但是正像

Bilgic 在文献[3]中提到的那样，现实世界中的社会网络实体属性和链接信息并不是完全已知的。假如当合著关系网络图中作者实体的类标号属性未知时，文献[3]中提出的算法就不能直接使用，这时就需要借助 ICCLP 来提高链接预测的性能。

正如前边所说的那样，当把与实体类标号属性有关的链接预测算法用到 ICCLP 中时，其链接预测的性能将会提高。这里把在文献[3]中提出的链接预测算法应用到 ICCLP 中，就可以实现同时对作者实体分类和预测合著关系。

在 ICCLP 算法中，当对作者实体类标号属性和合著链接信息完全未知的测试图执行一次集体分类算法(CCAlg)时，将得到作者实体的类标号属性。尽管不完全正确，这些类标号属性还是可以对链接预测提供帮助，因为它们可以帮助计算特征属性(是否至少有一个是多产的)的值。

如果一个作者是多产的，他很有可能在图中建立丰富的合著链接关系，与他合著的作者也更有可能通过他与他的合著者一起写文章。由此可以推得与多产的人合著的作者更有可能是多产的。由此可知合著关系可以对作者实体分类产生帮助，因此可以用集体分类算法来对合著关系网络图中的作者实体进行分类。

由上面两段分析可知。如果将文献[3]中提出的与实体类标号属性有关的链接预测算法用到 ICCLP 中，其集体分类部分和链接预测部分可以相互促进。正如文献[3]中的 ICCLP 算法那样迭代进行集体分类算法和链接预测算法，将会提高对于作者实体类标号属性未知的合著关系网络图进行链接预测的性能。

5 实验

本文用书目数据集：Elsevier BIOBASE (http://www.elsevier.com)作为实验的数据源。正像文献[9]中提到的那样，用 1998 年到 2002 年这 5 年的数据，其中前 4 年作为训练年份，最后 1 年作为测试年份。由于现实世界的社会网络中实体属性和链接信息并不是完全已知的，这里假设由 BIOBASE 构造的合著关系网络图中作者实体的类标号属性是未知的。这样的话，特征属性(是否至少有一个是多产的)不能被计算。在第 3 节提出的链接预测算法也就不能直接用来预测这个合著关系图。如果用这个链接预测算法，则不能用特征属性——是否至少有一个是多产的。实验表明用不包括特征属性(是否至少有一个是多产的)的第 3 节提出的算法来预测测试年份将出现哪些合著链接关系，预测的准确率仅有 87.12%。接着再用文献[9]中提出的算法处理相同问题。为了对比实验结果，也采用线性 SVM 作为分类算法，由文献[9]可知链接预测的准确率为 87.78%。

然后再用改进后的 ICCLP 算法来处理相同的问题。所谓改进的 ICCLP 算法，是指把第 3 节提出的链接预测算法用到 ICCLP 中。实验中，将迭代分类算法[12]的变体用作集体分类。用前 4 年的数据构造 ICCLP 算法中提到的训练图 G_{tr}。在训练图中，作者实体的类标号属性以及哪对作者将会在测试年份中合著都是已知的。在训练年份中没有合著的一对作者，如果在测试年份合著，同样构成他们之间的合著关系。在 ICCLP 算法中提到的测试图 G_{te} 中，作者实体的类标号属性及哪对作者将在测试年份合著都是未知的。由此可知，测试图中的合著链接数目比训练图中的少。

在上面提到的训练图上训练集体分类算法和链接预测算法，得到集体分类模型和链接预测模型。然后再像 ICCLP[3] 那样在测试图上迭代进行集体分类和链接预测。实验表明，这样进行链接预测得到的准确率为 88.34%。

结束语 实验表明，对于实体属性并不完全已知的社会网络，尤其是当实体的类标号属性未知时，可以利用改进的 ICCLP 算法提高链接预测的性能。所谓改进的 ICCLP 算法，是指将与实体的类标号属性有关的链接预测算法应用到 ICCLP 中。而且这样，ICCLP 的链接预测部分也比单一的链接预测性能更好。本文除了提出了一种用到实体类标号属性的链接预测方法外，还详细描述了合著关系网络图的结构。

尽管本文提出了一种与实体类标号属性有关的链接预测算法，并将其用到已有的 ICCLP 算法中以达到提高链接预测性能的目的，然而这种方法只能对合著关系网络图进行链接预测。由于实体的类标号属性与具体的社会网络有关，因此将来还需要研究适应于其它类型社会网络的类似链接预测方法。

参 考 文 献

[1] Han Jiawei, Kamber M. Data Mining: Concepts and Techniques. 北京：机械工业出版，2006

[2] Rattigan M, Jensen D. The case for anomalous link detection// 4th Multi-relational Data Mining Workshop, 11th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. 2005

[3] Bilgic M, Namata G-M, Getoor L. Combining Collective Classification and Link Prediction// Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007). 2007

[4] Huang Zan. Link Prediction Based on Graph Topology: The Predictive Value of the Generalized Clustering Coefficient LinkKDD2006. Philadelphia, Pennsylvania, USA, August 2006

[5] Popescul A, Ungar L H. Statistical relational learning for link prediction// IJCAI Workshop on Learning Statistical Models from Relational Data. 2003

[6] Yu K, Chu W, Yu S, et al. Stochastic relational models for discriminative link prediction. Advances in Neural Information Processing Systems, 2007

[7] Pavlov M, Pavlov R. Finding Experts by Link Prediction in Co-authorship Networks// expert finding iswc link network prediction workshop_fews. 2007

[8] Madadhain J O, Smyth P. Feature-based Link Prediction in Networks, Site Visit for UC Irvine KD-D Project. April 2004

[9] Hasan M-A, Chaoji V, Salem S, et al. Link Prediction Using Supervised Learning Link Analysis. Counterterrorism and Security, 2006

[10] Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery, 1998, 2(2): 1-47

[11] Platt J C. Fast training of support vector machines using sequential minimal optimization// Schölkopf B, Burges C J, Smola A J, eds. Advances in Kernel Methods: Support Vector Learning, Cambridge, MA: MIT Press, 1999: 185-208

[12] Lu Q, Getoor L. Link-based classification// Intl. Conf. on Machine Learning. 2003