基于用户查询的多关系群体挖掘改进算法

闫娜娜,刘 锋,李锡娟,耿 波
(安徽大学 计算机学院,安徽 合肥 230039)

摘 要:多关系群体挖掘是近年来快速发展的重要的数据挖掘领域之一。传统的群体挖掘方法是假定网络中只有一种关系,并且挖掘结果与用户需求无关。但现实中的社会网络中存在着多种关系。基于用户查询,不同的关系表现出不同的重要性。分析了多关系群体挖掘中关系提取的问题,提出一种新算法对满足用户期望的关系进行最优线性合并。利用获得的合并关系提高群体挖掘的精准性。

关键词:多关系网络;群体挖掘;关系提取
中图分类号:TP301.6 **文献标识码**:A **文章编号**:1673-629X(2008)06-0020-03

An Improved Algorithm of Community Mining from Multi-Relational Network Based on User Inquiry

YAN Na-na, LIU Feng, LI Xi-juan, GENG Bo
(Department of Computer Science and Engineering, University of Anhui, Hefei 230039, China)

Abstract: Community mining in multi-relational network is one of the major fields in data mining which develops fast in recent years. Most of the traditional methods on community mining assume that there is only one kind of relation in the network, and moreover, the mining results are independent of the users' needs. However, in reality, there exist multiple relationships in social network, and each of these may play a different role in a particular task. In this paper, analyze relation extraction of community mining from multi-relational network, and propose a new method for learning an optimal linear combination of these relations which can meet the user's expection. With the obtained relation, community mining accuracy can be improved.

Key words: multi-relational network; community mining; relation extraction

0 引 言

随着因特网和万维网的快速发展,网络群体和基于网络的社会网络繁荣起来。现存的大多数社会网络分析[1]都是假设只存在单一的社会网络,表现相关的、同质的关系,比如像网页链接[2]。但实际的社会网络是多关系网络,存在着多样的关系。每种关系又可单独处理成一个关系网络。这些关系在不同的任务中起着不同的作用。为发现拥有特定性质的群体,首先要识别出在群体中扮演重要角色的关系[3]。但这种关系并不是明确的存在的,因此要在关系网络中的群体挖掘之前发现这种隐藏关系。这个问题可以用数学建模为关系选择和提取。关系提取问题可以简单地定义如下:在多关系网络中基于标记样本(如由用户查询提供的),怎样评估不同关系的重要性,以及怎样得到现有关系的合并使之与标记样本的关系能最好的匹配。文中提出一种改进的关系选择和提取算法。基本思想就是用最优化问题对问题建模。用权值矩阵来表示每个关系特征。矩阵中的元素反映相应对象间的关系强度。此算法旨在找出权值矩阵的最优线性合并,可以更好地接近与标记样本相对应的权值矩阵,从而更好地满足用户需求。

收稿日期:2007-09-22
基金项目:安徽省自然基金项目(070412051);安徽高校省级重点自然科学研究项目(KJ2007A43)
作者简介:闫娜娜(1983-),女,安徽阜阳人,硕士研究生,研究方向为数据挖掘;刘 锋,教授,硕士生导师,研究方向为并行分布计算。

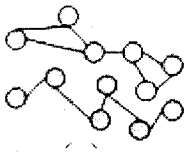
1 关系提取

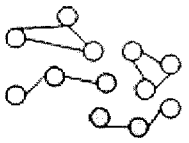
1.1 问 题

一个典型的多网络所包含的多种不同的关系,可用不同的图模拟。在群体挖掘问题中,这些图从不同的角度反映对象间的关系,提供出不同的群体。如图1所示的网络中可形成三种不同的关系。假设用户查询(一)要求的四个着色对象属于同一个群体,那么:

1)显然的,三种关系在反映用户信息需求上有不同的重要性。如图1所示,在反映用户信息需求中,关

系(a)最重要,关系(b)次之,关系(c)可被视为噪声。


关系(a)图


关系(b)图

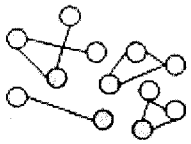
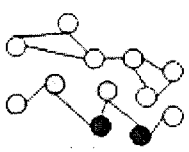
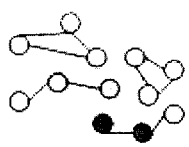

关系(c)图

图1　基于用户查询(一)的关系图

2)传统的社会网络分析(SNA)对这些关系不做区分。不同的关系被同样处理,且只是简单地结合在一起表述对象之间的结构关系。但是通过这个例子可见,关系(c)在关系合并中会起到消极作用。然而,若把关系依据其重要性进行结合,关系(c)就可以很容易地被舍弃,把关系(a)和关系(b)用于发现与用户需求一致的群体结构,就大大提高了群体挖掘的精确性。

与图1不同的是,在某些情况下用户可能提交更复杂的查询(如查询(二),图2给出一个例子。这个网络中的关系和图1是相同的,不同的是用户信息需求变了。用户要求着色轻的两个对象和着色重的两个对象要属于不同的两个群体。在这种情况下,三种关系的重要性也随之改变了。关系(b)变成最重要的,而关系(a)却变成消极的。


关系(a)图


关系(b)图

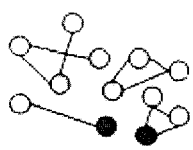

关系(c)图

图2　基于用户查询(二)的关系图

在多关系社会网络中,群体挖掘应依赖于用户提供的样本或信息需求。用户查询可以是复杂的。以前的群体挖掘技术只是基于单关系网络,而且和用户查询无关,不能处理像上述的复杂情形。

群体挖掘基于关系提取图,最有可能满足用户的信息需求。因此,文中着重研究多关系社会网络中的关系提取。关系提取可以是线性的也可以是非线性的。考虑非线性技术往往是不稳定的,可能出现过度适配问题,因此只考虑线性技术。

关系提取问题数学定义如下:

用图集 $G_i=(V,E_i)$ 表示对象集和关系集,$i=1,\cdots,n$,其中 n 是关系总数。V 是结点(对象)集,E_i 代表第 i 个关系中的边集。各边上的权值根据两对象间的关系强度定义[4]。用 $\boldsymbol{M}_i$ 表示与 G_i 相关的权值矩阵,$i=1,\cdots,n$。用图 $G^*=(V,E^*)$ 表示存在的隐藏关系,用 $\boldsymbol{M}^*$ 表示与 G^* 相对应的权值矩阵。给定一个标记对象集 $X=[x_1,\cdots,x_m]$,就对应有 $Y=[y_1,\cdots,y_m]$,其中 y_j 是 x_j 的类标(像这样的标记对象集本身也预示着隐藏关系 G^* 的局部信息),所要做的就是找出能够给出隐藏矩阵最好估计的这些权值矩阵 $\boldsymbol{M}^*$ 的线性合并。

1.2　基于回归的算法

算法的基本思想是找出一种合并关系使得群体之内的类相似度较高,同时群体之间的类相似度较低。

对每个关系,规定其最大强度(每条边的权值)为1。如下定义样本对象之间的目标关系矩阵:如果样本 i 和样本 j 类标相同,则 $\boldsymbol{M}_{ij}^+=1$;否则,$\boldsymbol{M}_{ij}^+=0$。其中 $\boldsymbol{M}^+$ 是 $m\times n$ 矩阵且 $\boldsymbol{M}_{ij}^+$ 表示样本 i,j 之间的关系。

有时,用户不能确定两个样本对象是否属于同一群体,只能确定两个对象属于一个群体的可能性。这种情况下,定义 $\boldsymbol{M}_{ij}^+$ 如下:

$$\boldsymbol{M}_{ij}^+=\text{prob}(x_i \text{ and } x_j \text{ belong to the same community})$$

引入向量 $\boldsymbol{a}=[a_1,a_2,\cdots,a_n]^{\mathrm{T}}\in R^n$ 来表示不同关系的合并系数。近似问题就可以特征化为解决下述的最优化问题:

$$\boldsymbol{a}^{\text{opt}}=\arg\min_{\boldsymbol{a}}\|\boldsymbol{M}^+-\sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{M}_i\|^2 \tag{1}$$

因为矩阵 $\boldsymbol{M}_{m\times m}$ 是对称的,可用一个 $m(m-1)/2$ 的空间向量 $\boldsymbol{V}$ 来代替。则函数(1)等价于

$$\boldsymbol{a}^{\text{opt}}=\arg\min_{\boldsymbol{a}}\|\boldsymbol{V}^+-\sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{V}_i\|^2 \tag{2}$$

函数(2)把关系提取问题建模成无约束线性回归问题。无约束线性回归的一个优点是它的解是递归闭合形式的,易于计算。然而线性回归问题的研究表明,无约束最小平方法并不是一个令人满意的解决方案,原因有以下两点:

(1)预测精确性。最小平方法方案估计经常有低偏差大变异[5]。而全局关系预测精确性有时要由收缩系数或把系数置0来提高。这样做要牺牲一点偏差来缩小预测关系强度的变异,从而提高全局关系预测的精确性。

(2)可解释性:聚类可能需要和特定的语义解释或应用相联系。根据大量的直接关系矩阵和相应的系数,可确定一个子集来展示强关系度。为了得到好的结果,应牺牲一些细节。

基于以上原因,应采用系数收缩技术。这样,对于每一个关系网络,规定边上的权值范围在[0,1]之间,且函数(2)中加入限制条件 $\sum_{i=1}^{n}\boldsymbol{a}_i^2\leq 1$。即转化为解决下述最小化问题。

$$\boldsymbol{a}^{\text{opt}}=\arg\min_{\boldsymbol{a}}\|\boldsymbol{V}^+-\sum_{i=1}^{n}\boldsymbol{a}_i\boldsymbol{V}_i\|^2 \tag{3}$$

其中$\sum_{i=1}^{n} a_i^2 \leq 1$

这种受限回归叫做岭回归。当自变量系统中存在多重相关性时，它可以提供一个比最小二乘法更为稳定的估计，并且回归系数的标准差也比最小二乘估计的要小。

2 实验验证

文中用 Iris 数据集验证此算法的有效性。Iris 数据集包含 150 条样本记录，分别取自三种不同类的鸢尾属植物的花朵样本，其中每条记录有 4 个特征属性：萼片长度 F1（sepal length），萼片宽度 F2(sepal width)，花瓣长度 F3(petal length)和花瓣宽度 F4(petal width)。其中 F3(petal - length)和 F4(petal - width)在挖掘隐藏的簇过程中尤其重要。

对每一个特征属性 Fr，构造对应的关系矩阵 $\boldsymbol{M}_{r,ij}$ 如下：

$$\boldsymbol{M}_{r,ij} = e^{-(x_i - x_j)^2} \quad (4)$$

Iris 数据集可以视为有 3 个隐藏群体的多关系社会网络。构造四个关系矩阵 $\boldsymbol{M}1, \boldsymbol{M}2, \boldsymbol{M}3, \boldsymbol{M}4$ 分别对应于四个特征属性。对 Iris 数据集进行可视化分析时，得到四个关系(特征属性)图，分别如图 3(a)，(b)，(c)，(d)所示，其中亮度代表两个对象间的关系强度。

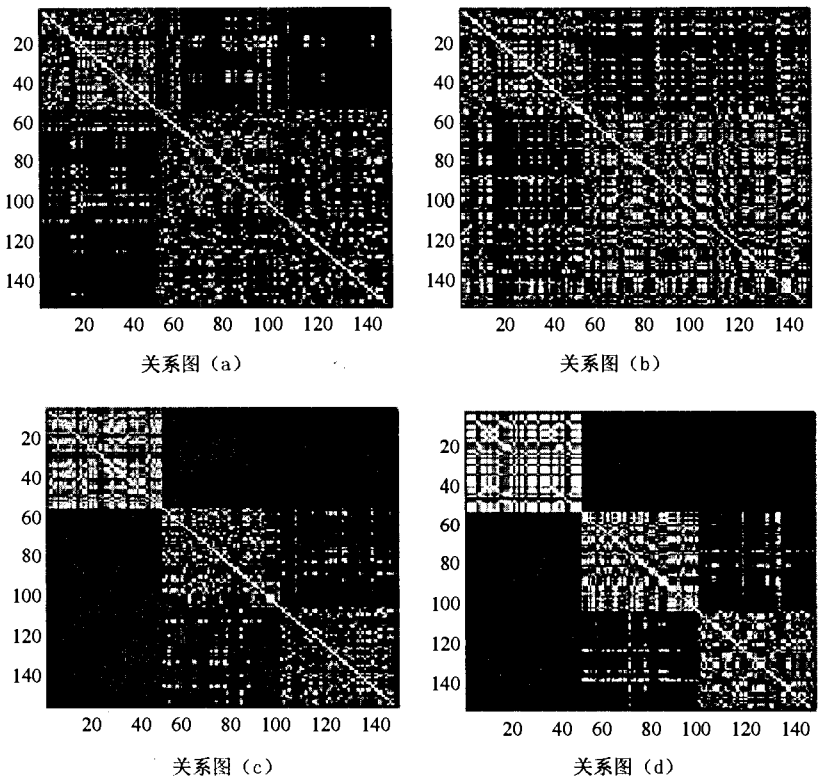


图 3 基于 Iris 数据集的模拟关系图

实验中，以给定的标记样本模拟用户查询，用 1.2 节中的基于回归的关系提取算法来提取关系。用标准化剪切算法[6]作为群体挖掘算法。群体挖掘结果的性能通过对比每个对象获得的群体标记与事实类标的结果来评估。

引入 r_i 表示对象 x_i 获得的群体标记，s_i 表示对象 x_i 的事实类标，则正确性 A 定义如下：

$$A = \sum_{i=1}^{n} \delta(s_i, \mathrm{map}(r_i))/n \quad (5)$$

其中，n 代表对象数，$\delta(x,y)$ 函数定义如下：如果 $x = y$，则 $\delta(x,y) = 1$；否则，$\delta(x,y) = 0$。排序映射函数 $\mathrm{map}(r_i)$ 是群体标记 r_i 到事实类标的等价映射。运用 Kuhn - Munkres 算法可找到最好的映射。提取关系定义如下：

$$\widetilde{\boldsymbol{M}} = \sum_{i=1}^{4} a_i \boldsymbol{M}_i \quad (6)$$

传统的群体挖掘算法不考虑用户提供的查询，把四个关系一样处理，即合并系数 $a_i = 0.25$，合并矩阵为 $\widetilde{\boldsymbol{M}} = \sum_{i=1}^{4} 0.25\boldsymbol{M}_i$。把此作为实验对比的基准关系。而当考虑用户查询，用文中的改进算法提取关系，四个关系的合并系数不相同。选取 10% 样本作为用户查询，进行关系提取，得到的四个合并系数为 0,0,0.5948,0.8039。得到的提取关系如图 4 所示。

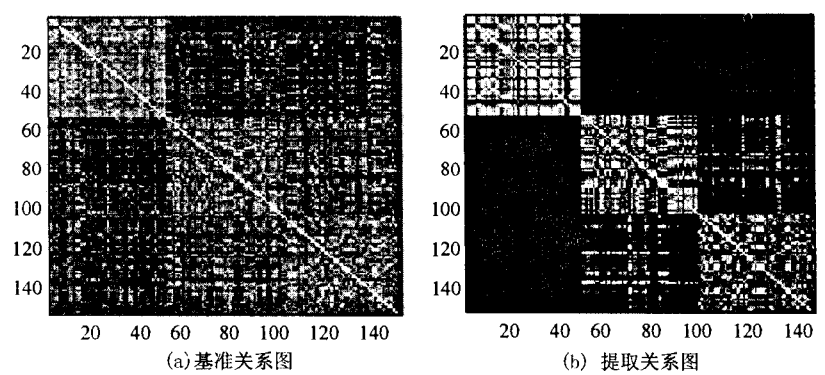


图 4 基准关系与提取关系图

实验再分别从每一类中随机选择 k 个样本作为用户查询，提取最优关系。应用标准化剪切算法进行群体挖掘中的关系提取，记录其正确性。此过程重复 100 次，取平均计算其性能。结果发现，用到的标记样本越多，算法的性能越好。基于不同样本率的群体挖掘的正

(下转第 27 页)

```
</dealWithPerson>
.....
</ProcessOfEngineeringDrawing>
```

图 5　工程设计蓝图文件

3　结束语

针对工作流程中的知识管理问题,提出了一种基于本体论的工作流知识管理系统架构,将工作流程文件的格式与结构建立在可交换与可重复使用的基础上,同时改善了词汇差异的问题,整合各领域事物间的关系与特性至工作流程中,借助领域性知识的明确定义,使得数据在传递与储存的过程中能正确无误地被处理。文中采用的本体论并非自动化的产生,在以后的研究中,可以考虑设计本体论的建构处理器,利用该处理器对领域的概念作分类,自动产生本体论,并采用文中转换本体论的方法,自动产生合适的 DTD。还可以通过本体论的表示法建立搜索处理器,搜索文件中的信息给用户或应用程序加以处理。

参考文献:

[1] Hollingsworth D. The Workflow Reference Model[EB/OL]. 2006. http://www.wfmc.org/standards/docs/tc003v11.pdf.

[2] Andersson B, Bider I, Perjons E. Integration of Business Process Support with Knowledge Management – A Practical Perspective[C]// Practical Aspects of Knowledge Management – 5th International Conference. Vienna, Austria: [s.n.],2004:227–238.

[3] Mou Yu-jie, Zhang Shen-sheng, Cao Jian. Providing knowledge support in business process: A context based approach [C]//Proceedings of IEEE International Conference on Systems, Man and Cybernetics. Hague, Netherlands: [s.n.], 2004:2143–2149.

[4] Chung P W H, Cheung L. Knowledge-based process management – An approach to handling adaptive workflow[J]. Knowledge-Based Systems, 2003, 16(3):149–160.

[5] Moore J, Inder R, Chung P, et al. Combining and Adapting Process Patterns for Flexible Workflow[C]//IEEE Database and Expert Systems Applications, Proceedings 11th International Workshop. London:[s.n.],2000:797–801.

[6] 沈兵虎,王　坚,潘瑞芳,等. 基于工作流技术的知识管理系统研究与设计[J]. 制造业自动化,2007,29(3):23–27.

[7] Erdmann M, Studer R. How to structure and access XML documents with ontologies[J]. Data & Knowledge Engineering, 2001, 36(3): 317–335.

(上接第 22 页)

确性如图 5 所示。

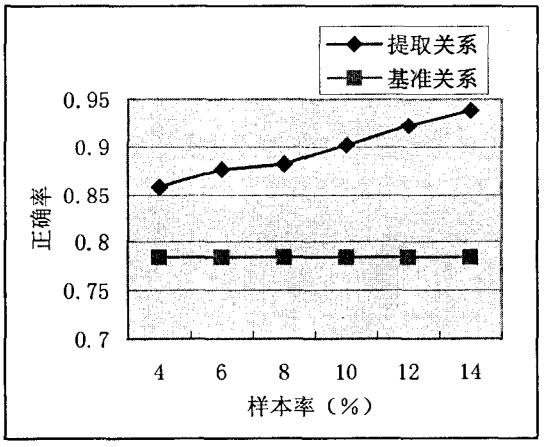


图 5　基于提取关系的群体挖掘的正确性

实验证明,改进的算法可提取基于用户查询(标记数据)的最优关系。利用标记信息,可优化关系提取。在多关系网络中,用户信息需求是多样的。这使得关系提取在社会网络分析的预处理过程中越来越重要。

3　结束语

社会网络中存在着多样、异构的关系。对这些关系进行有效的合并可以产生重要的更能适合用户信息需求的新关系。基于这样的想法,提出一种基于用户查询的新算法应用于关系提取。应用此算法进行关系提取和群体挖掘,可以获得精准的语义,提高群体挖掘效能。可以预见,这种依赖于查询的关系提取和群体挖掘在社会网络分析中将会引起许多潜在的新应用。

参考文献:

[1] Wasserman S, Faust K. Social Network Analysis: Methods and Applications [M]. Cambridge, UK: Cambridge University Press, 1994.

[2] Adamic L A, Adar E. Friends and neighbors on the web[J]. Social Networks, 2003, 25(3):211–230.

[3] Schwartz M F, Wood D C M. Discovering shared interests using graph analysis[J]. Communications of the ACM, 1993, 36: 78–89.

[4] Domingos P, Richardson M. Mining the network value of customers[C]//In: Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining. [s.l.]: ACM Press, 2001:57–66.

[5] Bjorck A. Numerical Methods for Least Squares Problems [M]. Philadelphia, PA: SIAM, 1996.

[6] Washio T, Motoda H. State of the art of graph-based data mining[J]. SIGKDD Explor. Newsl, 2003, 5(1):59–68.