DOI:10.7544/issn1000-1239.2016.20151172
53(4):776-784,2016

基于项目合作的社会关系网络构建

何贤芒[1,4] 陈银冬[2] 李 东[3] 郝艳妮[3]
[1](宁波大学信息科学与工程学院 浙江宁波 315211)
[2](汕头大学工学院 广东汕头 515063)
[3](国家自然科学基金委员会信息中心 北京 100085)
[4](复旦大学计算机科学技术学院 上海 200433)
(hexianmang@nbu.edu.cn)

A Construction for Social Network on the Basis of Project Cooperation

He Xianmang[1,4], Chen Yindong[2], Li Dong[3], and Hao Yanni[3]
[1](*Faculty of Information Science and Engineering*, *Ningbo University*, *Ningbo*, *Zhejiang* 315211)
[2](*College of Engineering*, *Shantou University*, *Shantou*, *Guangdong* 515063)
[3](*Information Center*, *National Natural Science Foundation of China*, *Beijing* 100085)
[4](*School of Computer Science*, *Fudan University*, *Shanghai* 200433)

Abstract For the time being, the social network based on paper cooperation has gained a great deal of attention, but there exists inaccurate entity recognition, failing to update data in time, and uncertain data quality etc. In view of this, this paper puts forward the cooperation on the basis of the history project application, and the problem of the entity recognition attributes to a clustering problem. The computational complexity of the problem is proved. Then the algorithm is proposed to settle the problem. Finally, the efficiency of the algorithm is verified by the experiments on real data.

Key words project cooperation; social network; entity recognition; clustering; computational geometry problems

摘 要 目前,基于论文合作关系的科学研究人员社会关系网络得到了极大的关注,但是存在实体识别不准确、数据更新不及时等数据质量问题.有鉴于此,提出利用历年项目申请书的合作关系,同时将实体识别问题归结为一个聚类问题,证明该问题的计算复杂度,然后提出了算法来解决该问题,最后在真实数据上验证算法的效率.

关键词 项目合作;社会关系网络;实体识别;聚类;计算几何问题

中图法分类号 TP311.131

收稿日期:2015-12-21;**修回日期**:2016-03-11

基金项目:国家自然科学基金项目(61103244,U1509213);广东省自然科学基金项目(2015A030313433);广东省高等学校优秀青年教师培养计划项目(Yq2013074);广东省普通高校特色创新项目(2015KTSCX036);广东省高校工程技术研究中心建设项目(GCZX-A1306);信息与通信工程浙江省重中之重学科开放基金项目;中国博士后科学基金项目(2013M540323);教育部人文社会科学研究项目(15YJA630069);汕头市科技计划项目(98)

This work was supported by the National Natural Science Foundation of China (61103244, U1509213), the Natural Science Foundation of Guangdong Province of China (2015A030313433), the Foundation for Distinguished Young Talents in Higher Education of Guangdong (Yq2013074), the Characteristic Innovation Project in Higher Education of Guangdong (2015KTSCX036), the Engineering and Technology Research Center of Guangdong Higher Education Institutes(GCZX-A1306), the Top Priority of the Discipline (Information and Communication Engineering) Open Foundation of Zhejiang, the China Postdoctoral Science Foundation (2013M540323), the Humanity and Social Science Foundation of Ministry of Education of China (15YJA630069), and the Shantou Science and Technology Foundation (98).

通信作者:陈银冬(ydchen@stu.edu.cn)

由于社交网络的繁荣发展和广泛应用,越来越多的研究者将其科学研究和应用开发的注意力集中到社会网络这种虚拟世界当中.社会网络分析已然成为社会学、地理学、经济学、信息科学等诸多学科的重要研究内容,其研究涉及了数据挖掘、知识管理、数据可视化、统计分析、社会资本、小世界理论、信息传播等多个学科.基于社会关系网络数据进行数据挖掘和潜在模式分析比传统数据统计分析更加科学、效果更好、应用前景更突出,通过对社会网络进行挖掘可以获得比简单实体数据更详实(如实体在社会网络中的关系)、更准确(如挖掘实体间的关系)的信息.

目前国内已有的基础研究人员社会关系网络主要是基于论文合作关系的信息服务平台,具有代表性的工作有深圳爱瑞斯公司研发的科研之友[1]、清华大学知识工程库开发的 ArnetMiner[2]、CCF 学术空间(仅限计算机领域的合作关系)[3]、中国人民大学研发的学术空间 ScholarSpace[4]等.这些平台共同的特点是提供了针对网络中海量文献检索服务,同时从作者、期刊/会议、工作单位、学科领域、合作者关系等多个角度进行统计与分析,还挖掘出基于论文合作的合作关系.但是上述平台的数据主要抽取网络中的不确定性数据,存在数据来源多样化、格式不一致、数据不准确、更新不及时等数据质量问题,导致实体识别不准确.本文的研究动机有如下 2 个:

1) 张冠李戴

在实践中,作者中文姓名重复的现象非常普遍.比如中文名"张伟",我们对 CNKI 数据库经过专门的实体识别后发现,存在 460 多个张伟.对这众多似是而非模棱两可的"张伟",哪怕采用人工尚有不小难度,更何况是计算机呢?"张伟现象"的根本原因在于从论文中提取的信息太少,只包括姓名、单位、邮箱、期刊名、论文共同作者等有限信息,而且在关于邮箱信息上往往也只有通信作者的邮箱(如果考虑到格式问题,很多作者的邮箱也不完整).因此,基于论文合作的社会关系网络的构建,即使考虑领域信息和共同合作关系,也无法做到准确的识别.

"张伟现象"仅仅是以论文合作社会关系网络的构建中的一个普通案例.事实上,我们在尝试构建 3 000 多位国家杰出青年基金获得者的社会关系网络时,依然存在识别准确性不高的现象,特别是论文与作者间张冠李戴的现象相当突出.因此,基于论文合作的社会关系网络难以做到对实体的准确识别.

2) 项目合作

论文和项目都是科研工作者最主要的科研成果形式,也自然是科研合作最常见的表现形式.既然基于论文合作构建社会关系网络比较困难,那么是否可以基于项目合作来构建社会关系网络呢?通过项目共同参与申请,将实体联系成一个复杂的社会关系网络.在社会关系网络中,每一个实体都是一个顶点,同一申请书中的申请人 A 与参与者 B 之间存在有向边:$B\rightarrow A$,其他参与者之间没有边关联.由于申请人也往往是其他项目中的参与者,如此便构成了一个复杂的社会关系网络.

以国家自然科学基金委员会(简称基金委)每年接收的申请书为例,目前已累积 1000 多万条申请人与参与人的信息,这些申请者的信息包括姓名、性别、单位、邮箱、身份证号、出生日期、职称、申请人/参与者等.通过对真实数据分析发现,绝大多数信息都是完整准确的.需要强调的是:同一实体在不同申请书中的信息有可能并非完全一致.比如由于工作调动而导致单位、邮箱等信息的变化.而更加特别的现象是,部分实体基于某些原因(比如规避基金委的申请限项规定)而故意错误填写某些关键信息(比如身份证号码).表 1 给出了项目申请书合作关系的一个示例.其中,id,prp_code,name,sex,org_name,

Table 1 An Example of Cooperation in Project Application

表 1 项目申请书合作关系示例

Id	Prp_code	Name	Sex	Org_name	Birthday	Email	Is_pc	Card_code
1	11030007	Zhang Wei	Male	Wuhan University	1977-02-01	zhangw@163.com	1	GT23697
2	11030007	Li Qiang	Male	WHU	1968-01-01		0	000000123456789012
3	11330211	Li Qiang	Male	Wuhan University	1968-01-01		0	000000123456789012
4	11572001		Male	Hangzhou	1977-02-01	zhangw@163.com	0	GT23697
5	10601094	Zhang Wei	Male	Wuhan University		zhangw@163.com	0	gt3697
6	10601094	Li Qiang	Male	Wuhan University	1968-01-01		0	000000123456789012
7	10501320	Li Qiang	Male	Wuhan University			0	
8	10501320	Zhang Wei	Male	Wuhan University	1977-02-01	zhangw@163.com	1	gt23697

birthday,email,is_pc,card_code 等字段分别表示元组序号、项目受理号、姓名、性别、单位、出生日期、邮箱、是否项目负责人、身份证号等. 基于隐私保护考虑,已对身份证号进行了替换处理. 在通过手工整理与实体识别处理后发现:基于项目合作的社会关系网络关系准确、可靠性高,为科学基金项目管理提供辅助检索与查询,保证科学基金的公正与公平,具有重要的研究价值和现实需求.

本文的主要工作包括 3 个方面:

1) 提出了一个基于项目合作的社会关系网络的构建设想,并从中发现一个重要现象:3 个属性值相同的元组可以认定为同一实体,从而引出本文讨论的主要问题.

2) 研究求解该问题的计算复杂度. 通过证明该问题等价于计算几何中的 USEC 问题,说明该问题的计算复杂度的下界满足 $\Omega(n^{4/3})$,除非 USEC 问题和 Hopcroft 问题在算法理论上有重大突破.

3) 利用数据聚类方法构建一个基于项目合作的社会关系网络,并提出高效的解决算法,同时基于真实数据对算法运行效率进行了验证.

1 基本概念与问题定义

给定一个数据集 $T(A_1,A_2,\cdots,A_d)$,假定 T 共有 d 个属性,用 $t[A_i]$ 表示其属性值,数据集中的一个点称为对象.

定义 1. ε-邻域 $B(p,\varepsilon)$. 给定对象 p,其半径 ε 内的区域称为该对象的 ε-邻域.

定义 2. 核心对象(core point). 如果给定对象 ε-邻域 $B(p,\varepsilon)$ 内的样本点数大于等于 $MinPts$,则称该对象为核心对象,其中参数 $MinPts$ 是一个预先给定的正整数.

定义 3. 密度可达. 对于数据集 T,如果存在一个对象链 $p_1=q,p_2,p_3,\cdots,p_{k-1},p_k=p$,其中 $p_1,p_2,\cdots,p_{k-1}$ 是核心对象,且 $p_{i+1}\in B(p_i,\varepsilon)$,$i=1,2,\cdots,k-1$,则称 q 密度可达 p.

例 1. 图 1 给出一个密度可达的示例. 此时 $MinPts=3$,P_7 密度可达 P_5,对象链是 P_7,P_6,P_5,但 P_5 并非核心对象.

在数据整理的实践过程中,我们发现了重要现象(以下统称为“三属性现象”):

观察 1. 申请人数据库中,在性别相同的情况下,任何 2 条元组只要在姓名、单位、邮箱、身份证、出生日期等 5 个主要属性上有 3 个以上的值相同,

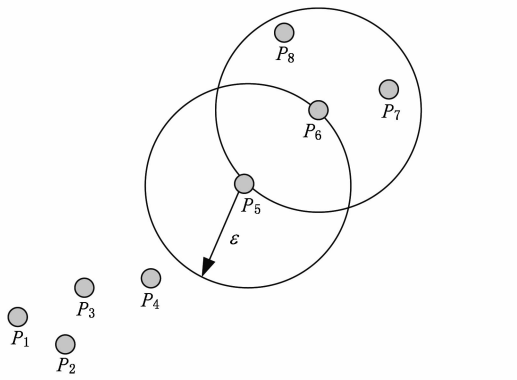


Fig. 1 An example for $MinPts=3$.

图 1 $MinPts=3$ 的示例图

可认定为同一实体. 根据“三属性现象”,在表 1 中,元组 t_1,t_5,t_8,t_4 可认定为同一实体,元组 t_2,t_3,t_6 也可认定为同一实体,而对元组 t_7 则无法确定其是否与元组 t_2,t_3,t_6 为同一实体.

此外,对于一些明显错误的证件号码比如 123456,2222222,00000000 等进行去除操作,同时注意到 2 个不同的申请人伪造了相同的证件号码可能性是很低的,比如表 1 中的证件号码 000000123456789012,尽管是错误的,但仍然有参考价值.

定义 4. 3-属性值相同(3-attribute value same). 任取数据集 T 中 2 条元组 $t_1,t_2\in T$,如果存在至少 3 个属性值相同,便认定为同一实体.

定义 5. 距离. 2 个元组 t_1,t_2 的距离 $dist(t_1,t_2)$ 定义如下:

$$dist(t_1,t_2)=\begin{cases}1, & |i:1\leqslant i\leqslant d,t_1[A_i]=t_2[A_i]|\geqslant 3;\\ 0, & |i:1\leqslant i\leqslant d,t_1[A_i]=t_2[A_i]|<3.\end{cases}$$

即若元组间取值相等的属性数量大于等于 3,则距离为 1,否则距离为 0.

定义 6. 类(cluster). 类(cluster)C 是数据集 T 的非空子集,且满足条件:

1) 3-属性相同. 任取元组 $t_1,t_2\in C$,一定存在对象链 $p_1,p_2,\cdots,p_{k-1},p_k\in C$,其中 $p_1=t_1,p_k=t_2$,满足距离 $dist(p_i,p_{i+1})=1,i=1,2,\cdots,k-1$.

2) 极大性. 任取 $t_3\in T-C$ 和 $t_1\in C$,则 $dist(t_3,t_1)=0$.

例 2. 表 1 中 $dist(t_1,t_5)=1,dist(t_5,t_8)=1$,虽然 $dist(t_1,t_4)=0$,但由于 $dist(t_8,t_4)=1$,因此 $C_1=\{t_1,t_4,t_8,t_5\}$ 构成一个类,表 1 最终分成 3 个类 $C_1=\{t_1,t_4,t_8,t_5\},C_2=\{t_2,t_3,t_6\},C_3=\{t_7\}$.

在完成聚类与类的定义之后,现在给出本文要解决的主要问题.

定义 7. 问题 1. 给定一个数据集 T 和 d 个属性

$T(A_1, A_2, \cdots, A_d)$,$d>3$,按照上述类和距离的定义,将数据集 T 分成不同的类.

2 聚类问题的计算复杂度

首先给出 2 个计算几何问题和若干已经证明的相关结论.

2.1 2 个计算几何问题

定义 8. USEC 问题. 给定一些半径 r 相同的球集合 S_{Ball} 和点集 S_{pt},判断是否存在某个点 $p\in S_{\text{pt}}$,p 在某个球内. 其中,点和球均为 d 维数据,d 是常数.

定义 9. Hopcroft 问题. 给定 2-D 空间上的点集 S_{pt} 和一些线 S_{Line},判断是否存在某些点 $p\in S_{\text{pt}}$,p 在 S_{Line} 的某些线上.

定义 10. Hopcroft 困难问题[5]. 如果问题 X 可以在 $O(n^{4/3})$ 内解决,表明存在算法在 $O(n^{4/3})$ 内解决 Hopcroft 问题,那么问题 X 是 Hopcroft 困难问题. 换句话说,如果 Hopcroft 问题能在 $\Omega(n^{4/3})$ 时间内解决,那么问题 X 也存在同样时间复杂度的算法.

设 $n=|S_{\text{pt}}|+|S_{\text{Ball}}|$,当 $d=3$ 时,USEC 问题可以在 $O((n\log n)^{4/3})$ 内解决,而寻找时间复杂度 $O(n^{4/3})$ 的算法依然是公开难问题. 类似地,设 $n=|S_{\text{pt}}|+|S_{\text{Line}}|$,研究认为 Hopcroft 问题的时间复杂度是 $\Omega(n^{4/3})$,而对于 USEC 问题则有结果:

引理 1[6]. 当 $d\geqslant 5$ 时,USEC 问题是 Hopcroft 难问题.

2 个计算几何问题的示意图如图 2 所示:

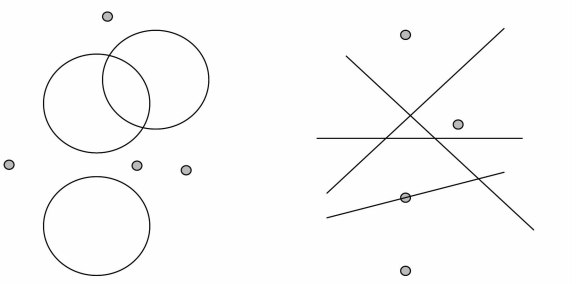


Fig. 2 Two computational geometry problems.

图 2 2 个计算几何问题

2.2 计算复杂度

下面证明问题 1 与 USEC 问题等价,进而说明其与 DBSCAN 问题等价.

定理 1. 当 $d\geqslant 4$ 时,问题 1 与 USEC 问题等价.

证明. 对于任意维数 $d\geqslant 4$,如果我们能在 $f(n)$ 时间内解决问题 1,那么就可以在 $f(n)+O(n)$ 时间内解决 USEC 问题. 考虑到 USEC 问题是定义 $\mathbb{R}^d$ 空间上的 S_{pt} 点集和半径相同的 S_{Ball} 球集合. 设算法 A 能在 $f(n)$ 内完成问题 1. 我们设计一个算法 1 在 $f(n)+O(n)$ 时间内回答 USEC 问题,$n=|S_{\text{pt}}|+|S_{\text{Ball}}|$. 证毕.

算法 1. USEC 问题回答算法.

① 记 S_{Ball} 的球心集合为 S_{center},设数据集 $P=S_{\text{pt}}\cup S_{\text{center}}$,球半径 $\varepsilon=1$;

② 在数据集上运行算法 A;

③ 如果存在某个点 $p\in S_{\text{pt}}$ 和 $p'\in S_{\text{center}}$ 在同一个类中,那么回答 Yes;

④ 否则回答 No.

显然,算法 1 的执行时间不超过 $f(n)+O(n)$,下面分 2 种情形证明其正确性.

情形 1. 回答 Yes 情形.

首先注意到如果把一个类里面的点作为顶点,若顶点间距离为 1 则用边连接,那么在同一个类中的点便构成一个连通图. 若 p,p' 处于同一个类中,则存在一个对象链:$p_1=p,p_2,p_3,\cdots,p_{k-1},p_k=p'$,从而,一定存在 p_i,p_{i+1} 使得 $p_i\in S_{\text{pt}}$,$p_{i+1}\in S_{\text{center}}$. 考虑到 $p_i\in B(p_{i+1},\varepsilon)$,从而球覆盖了点 p_i.

情形 2. 回答 No 情形.

相反地,如果 p 和 p' 并非处于同一类中,那么二者之间的距离为 0,故不存在某个 p_i 属于 p' 的 ε-邻域 $B(p,\varepsilon)$,因此回答是 No.

定理 2. 当 $d\geqslant 4$ 时,问题 1 与 BSCAN 问题等价.

Tao 等人[7]证明了 DBSCAN 算法与 USEC 问题定价. 结合定理 1,定理 2 的结论不言而喻. 事实上,问题 1 与 DBSCAN 问题的等价性是明显的,只需巧妙设置 DBSCAN 问题中的 2 个参数($Minpts=1,\varepsilon=1$)即可. 此时,$Minpts=1$ 意味着 DBSCAN 里面所有点都是核心对象.

综上,本文得到问题 1 的计算复杂度的结论如下:

① 当 $d\geqslant 4$ 时,问题 1 与 USEC 问题和 DBSCAN 算法均等价.

② 当 $d=4$ 时,问题 1 需要在 $\Omega(n^{4/3})$ 时间内解决,除非 USEC 问题可以在 $O(n^{4/3})$ 解决.

③ 当 $d\geqslant 5$ 时,问题 1 是 Hopcroft 难问题,亦即问题 1 需要在 $\Omega(n^{4/3})$ 时间内解决,除非 USEC 问题可以在 $O(n^{4/3})$ 内解决.

3 相关工作

目前社会关系网络研究中最受关注的是,除了基于论文合作的社会关系网络外,还有基于论文

引用的社会关系网络[8]、基于电话的社会关系网络[9]、基于关注的社会关系网络(比如 Twitter、微博)[10]、基于专利合作的社会关系网络[11]等. 同时,也有一些相应研究与具体应用. 现在,已有超过 20 多个社会关系数据集发布于网上供公众研究与测试.

对于社会关系网络的研究分析主要包括 3 个方面:

1) 社会影响分析与行为分析. 主要包括网络上节点与边的影响力分析、节点与边的度量、节点介数等[12];Kumar[13]研究了网站上博客信息动态传递行为、应用传染病机制来对主题传播行为进行建模. Gruhl 等人[14]探索了社会网络会话结构的形成,并用一种数学模型来刻画用户对话行为. 文献[15]研究了社会关系网络中强连接节点问题. Liben-Nowell 和 Kleinberg[16]探讨了人与人之间信息传播过程来重构大规模分发互联网连锁信的传播机制,他们发现连锁信通过一种很窄且很深像树一样的模式传播. Yang 等人[17]分析了 Twitter 上用户转发行为,并提出了一个半监督的框架来预测用户的行为; Myers 等人[18]研究了消息的扩散过程是如何受外来消息源的影响;文献[19]提出了一个 SPIKEM 模型来表示消息传递的上升与下降模式. Huang[20]研究了社会关系网络中的群体形成问题尤其是 triad 形成问题,并建立一个模型来在线预测动态社会关系网络中 closed triad 的形成.

2) 社会关系分析. Tang 等人[21]研究了异构网络环境下如何区分不同类型的社会关系,提出了一个整合了社会关系理论框架,可以有效地提升社会关系类型的区分. Wang 等人[22]提出了一种挖掘导师与学生关系的 TPFG 模型,该模型以论文合作的社会关系作为研究对象. 文献[23]提出了一个监督随机游走算法来估计关系中的强度,Diehl 等人[24]提出了一个排名函数来区别上下级关系. 文献[25]提出了移动通话数据中的几种关系模式,并利用这些模式来推断朋友关系网络. Sun 等人[26]研究了多个类型动态网络中社区演化.

3) 社会网络结构分析. Yang 等人[17]研究了社会关系网络的用户转发行为;Zhang 等人[27]发现社会关系网络的局部性,也就是一个人容易受到其最亲密的朋友影响,形式化表述了这个现象并建立数学模型来预测用户的转发行为. 结构洞(structural holes)理论[28]也在社会关系网络结构分析中得到了应用,文献[29]研究了社会关系网络的结构洞形成问题,Lou 等人[30]描述了如何挖掘大型社会关系网络的结构洞中 top-k 问题,并证明了该问题是 NP-hard.

此外,关于 DBSCAN 算法的研究,最早是由 Ester, Kriegel, Sander 和 Xu 在 KDD1996 会议[31]上提出,一经提出就受到了极大的关注. 值得一提的是,DBSCAN 算法的提出者一直误以为其时间复杂度为 $O(n\log n)$,直到 2013 年才由 Gunawan [32]指出其实际时间复杂度是 $O(n^2)$,陶宇飞等人[7]在 2015 年 SIGMOD 会议上证明了 DBSCAN 算法的时间复杂度:当数据维度是 $d=3$ 或 $d=4$ 时,DBSCAN 算法与 USEC 问题等价;当 $d\geqslant 5$ 时,是 Hopcroft 难问题;同时,他们还提出了一个时间复杂度为 $O(n)$ 的近似算法. 特别地,对于 2 维数据,Han 等人[33]提出了时间复杂度为 $O(n\log n)$ 的算法,同时证明了类的数量正好等于图的连通分量的数量. 此外,对于 Hopcroft 问题,目前最好算法的时间复杂度大致比 $O(n^{4/3})$ 稍大,具体可参阅文献[34].

4 实体识别算法

在本节,我们将提出实体识别算法,即问题 1 的求解算法. 由第 2 节的讨论可知,问题 1 的计算复杂度下界是 $\Omega(n^{4/3})$,除非 USEC 问题与 Hopcroft 问题有了重大的理论突破,因此直接设计时间复杂度为 $O(n\log n)$ 不太现实. 此外,注意到近似算法对问题 1 没有特别的意义,因此本节算法的设计是精确(exact)算法.

根据观察得出的结论,对数据集 T 采用交叉比对的方法进行实体识别. 算法 2 给出了一个初等的 Naïve 算法,该算法的时间复杂度 $O(n^2)$,n 为数据集 T 的大小. 4.2 节将提出一个改进算法.

4.1 基于规则的 Naïve 算法

算法 2. 基于规则的 Naïve 交叉比对算法.

输入:数据集 T;

输出:类 $C_1, C_2, \cdots, C_m$.

① for $i=1$ to $|T|$
② for $j=i+1$ to $|T|$
③ if ($dist(t_i, t_j)=1$) /* 如果 3 个属性值相同 */
④ 将 t_i 和 t_j 标记为同一个类;
⑤ end if
⑥ end for
⑦ end for
⑧ 按顺序输出每个类及其所包含的元组.

算法 2 的直接想法是检查每对元组是否满足 3-属性值相同,若相同就标记为同一实体,并放入同一类中. 算法 2 虽然复杂度较高,但算法简单,可作为基准算法和检验后续算法的准确性.

4.2 基于分组的改进识别算法

对于大数据集,时间复杂度为 $O(n^2)$ 的算法 2 几乎是不可接受的,因此需要改进算法.问题 1 的本质是找出所有至少 3 个属性值相同的类 $\{C_1, C_2, \cdots, C_m\}$.我们可以先找出某一个属性值相同的集合 $\{T_1, T_2, \cdots, T_k\}$,然后对每个 T_i 进行交叉比对,从中找出并确定类 $\{C_{11}, C_{12}, \cdots, C_{1s}\}$.以上便是算法 3 的基本思想.

为了解决问题 1,算法 3 至多运行 $d-2$ 次,记录每次算法的结果 $\{C_{i1}, C_{i2}, \cdots, C_{is}\}$,设元组 t 每次落在不同的类 $C_{1i}, C_{2j}, \cdots, C_{(d-2)k}$,将这些类合并为一个类,当所有的元组都完成合并时,算法便终止.

算法 3. 基于分组的识别算法.

输入:数据集 T、T 的某属性 $Attr$;

输出:类 $C_{11}, C_{12}, \cdots, C_{1m}$.

① 以属性 $Attr$ 为主进行排序;

② 将 T 划分为 $T_1, T_2, \cdots, T_k$,使得 $T=T_1 \cup T_2 \cup \cdots \cup T_k$,$T_i \cap T_j = \varnothing$;

③ 对每个 T_i 进行交叉比对;

④ for $i=1$ to $|T_i|$

⑤ for $j=i+1$ to $|T_i|$

⑥ if $(dist(t_i, t_j)=1)$

⑦ 将 t_i 和 t_j 标记为同一个类;

⑧ end if

⑨ end for

⑩ end for

⑪ 按顺序输出每个类及其所包含的元组.

算法 4 是针对上述"合并"操作的算法,其基本思想是对于每个元组 t,找出其落在 2 次不同分组的类 C_1, C_2,计算差集 C_2-C_1.如果差集 C_1-C_2 是空集,说明元组 t 前后 2 次不同的聚类后的结果相同,此时不需要进行合并;如果差集 C_1-C_2 非空,那么说明元组 t 前后 2 次不同的聚类分在不同的类中,因此需要将类 C_1, C_2 合并.

算法 3 的时间复杂度等于 $O(n\log n+\sum_{i=1}^{k}|T_i|^2)$,因此,其算法的效率主要依赖于属性 A_t 分组 T_i 的大小.如果 T_i 都比较小,其时间复杂近似等于 $O(n\log n)$;如果某个 T_i 比较大,那么近似等于 n^2,因此算法 2 的时间复杂度下界是 $\Omega(n\log n)$.合并算法的时间复杂度上界是 $\sum_{i=1}^{n}(\sum_{j=1}^{d-3}C_{ij}\log C_{ij})$.

算法 4. 合并算法.

输入:2 个聚类结果 l_1 和 l_2;

输出:分组集合 l_1.

① for $i=1$ to $|T|$

② 设 t_i 在 2 次聚类 l_1 和 l_2 的分组分别是 C_1 和 C_2;

③ $C_i=C_2-C_1$; /* 2 个分组的集合差集 */

④ if $C_i=\varnothing$ /* 前后 2 次聚类结果相同 */

⑤ continue;

⑥ else /* 前后 2 次的聚类结果不同,需要合并 */

⑦ 对 C_i 中的每个元组 t

⑧ $merge(C_1, C_j)$ /* C_j 是 t 在 l_1 中所在的类 */

⑨ end if

⑩ end for

5 数据实验报告

5.1 数据预处理

在申请人的数据库中,发现有些申请书的某些数据是故意填写错误,比如不存在的证件号码、依托单位信息变更、不符合规则的电子邮件等.预处理主要任务是去除这些不规范(甚至错误)的申请人数据.具体做法是人工筛选出明确错误的数据共 30 621 条,直接将其从数据测试集中去掉.

用于测试的数据集是从国家基金委 1986—2014 年的申请书中提取的合作关系,共计 9840477 条数据,其中性别为男和女的 2 个数据集各有 6 092 060 条和 3 748 417 条.每条元组含 8 个属性(项目编号、姓名、单位、邮箱、性别、是否主持人、证件号码、出生日期).为了防止信息的泄露,所有的数据通过 AES 加密,相同的数据加密后依然相同,因此加密后的数据对算法没有影响.为了测试算法的可扩展性,又分别从男与女 2 个数据集中随机选取了 1 万、5 万、10 万、20 万、50 万、100 万、200 万和全部的数据作为数据集分别测试,分别标记 M1w,M5w,M10w,M20w,M50w,M100w,M200w,MTotal 和 F1w,F5w,F10w,F20w,F50w,F100w,F200w,FTotal.

下面将 Naïve 算法(标记为 NA)与改进算法(标记为 PB,包括算法 3 和算法 4)进行比较,2 个算法均在相同的数据集上进行以保证算法的可比性.实验主要目的是验证算法的正确性与算法的效率.算法正确性通过比较 2 个算法识别出来的实体结果来验证,算法效率通过比较算法的运行时间来验证.本文的实验环境为 Intel® Xeon® CPU 2.4 GHz 和 RAM 8 GB,所有的算法和评估程序均采用 C++ 实现.

5.2 算法结果

表 2 和表 3 分别给出了实体识别后的实体数量,实验结果表明 2 个算法出来的结果是完全一致的.从表 2 和表 3 可以看出,从 1986 年至今我国从事自然科学研究的学者共有 2 862 749 人,其中女性与男性分别有 1 171 655 和 1 690 394 位,包括了各种行业各个级别的自然科学研究者,比如教授、副教授、教授级高级工程师等高级职称研究者和博士生、研究生等各个层次.

Table 2 Number of Entities (Female)

表 2 实体数量(女研究者)

Data Set	Number
F1w	9 011
F5w	39 992
F10w	67 257
F20w	120 869
F50w	249 752
F100w	389 162
F200w	603 019
FTotal	1 171 655

Table 3 Number of Entities (Male)

表 3 实体数量(男研究者)

Data Set	Number
M1w	8 744
M5w	40 006
M10w	67 345
M20w	120 997
M50w	250 088
M100w	390 347
M200w	609 994
MTotal	1 690 394

5.3 运行时间

表 4 比较 Naïve 算法(NA)与改进算法(PB)的运行时间.从运行时间上看,NA 算法的运行时间明显比 PB 算法要短,而且 NA 算法的运行时间特征明显:运行时间与数据集大小的平方成正比,比如当数据集大小为 50 万,运行时间是 7 776 s 左右,从而可以得出在 100 万数据集上的运行时间大概是 7 776 s 的 4 倍,与实际运行时间相处无几.PB 算法共需要进行 3 次数据集的划分,分别基于姓名、出生日期与证件号码进行划分后合并.从时间上来看,2 个算法时间差别比较显著,尤其是当算法在女研究者数据全集上进行时 NA 算法执行 112 h,如果在男研究者数据全集上运行预计需要300 h,因此不在男研究者全集中执行 Naïve 算法.

Table 4 Running Time

表 4 运行时间 s

Data Set	NA	PB
F1w	2.84	0.032
F5w	74.38	0.281
F10w	316.30	0.608
F20w	1 317.30	1.669
F50w	7 776.20	5.834
F100w	30 089.00	1 465.000
F200w	121 012.00	1 584.000
FTotal	421 097.00	23 011.000

从 PB 算法来看,基于不同的属性划分会产生不同的计算代价,其主要原因跟属性值的分布密切相关,如果某个属性在某个值上的重复值较多,那么交叉验证时间比较多.表 5 中 PB-1 是算法 3 选择姓名、出生日期与证件号码进行了 3 次划分,而 PB-2 选择姓名、邮件和证件号码进行了 3 次划分,二者时间差别很大,这是由于 Email 属性值上重复值不多,而出生日期重复的情况比较多.表 6 比较 2 个算法在男研究者数据集上的运行时间,结果是完全类似的.

Table 5 Running Time of Different Attributes (Female)

表 5 不同的属性划分比较(女研究者) s

Data Set	PB-1	PB-2
F1w	0.032	0.062
F5w	0.281	0.250
F10w	0.546	0.593
F20w	1.669	1.373
F50w	5.834	16.114
F100w	1 465.000	33.805
F200w	1 584.000	116.486
FTotal	23 011.000	504.630

Table 6 Running Time of Different Attributes (Male)

表 6 不同的属性划分比较(男研究者) s

Data Set	PB-1	PB-2
M1w	0.062	0.078
M5w	0.250	0.281
M10w	0.624	0.609
M20w	1.497	1.389
M50w	4.914	26.567
M100w	1 049.000	33.650
M200w	1 099.000	435.530
MTotal	55 826.000	2 613.000

表7给出了PB-1选择姓名、出生日期与证件号码作为分组的属性,而PB-2选择姓名、邮件和证件号码作为分组的属性,算法的主要时间是3次分组和2次合并(表7中分别记为PAR1,PAR2,PAR3,MER1和MER2).从表7可以看出,算法的主要运行时间在第2次分组上,而选择邮件作为分组的依据比选择出生日期节省较多的时间,这是由于邮件重复的情形比出生日期要少很多.

Table 7 Composition of Running Time

表7 运行时间的构成 s

Data Set	PB-1	PAR1	PAR2	PAR3	MER1	MER2	PB-2	PAR1	PAR2	PAR3	MER1	MER2
M100w	1049	1.03	**1039**	1.0	0.46	0.43	33.65	1.1	**24**	1.0	0.45	0.43
M200w	1099	3.00	**1074**	3.8	1.00	0.95	435.53	2.9	**410**	3.9	1.00	1.00
MTotal	55826	16.00	**55646**	98.0	3.60	3.20	2613.00	15.0	**2414**	116.0	3.70	3.50

6 总 结

如何从海量数据中发现基础研究人员的合作关系,对于科学基金项目的全过程管理,为科学基金项目管理提供辅助检索与查询,保证科学基金的公正与公平,具有重要的研究价值和现实需求.针对历年累积的申请书中的合作关系,本文提出了基于项目合作关系的基础研究人员社会网络关系的构建.从实验结果来看,1986—2014年我国从事自然科学研究的工作者大概有280多万,其中男女比例大概是1.44:1.针对项目合作社会关系网络研究,我们将在社会网络结构分析、社会关系分析、社会影响分析与行为分析等方面开展工作,具体而言,包括基于项目合作网络的结构特征,从项目合作中挖掘出导师与学生关系、3人组(triad)形成等,这些工作的开展将会从另一视觉去审视基础研究人员的社会关系网络.

参 考 文 献

[1] Liu X, Guo Z, Lin Z. A local social network approach for research management [J]. Decision Support Systems, 2013, 56(12): 427-438

[2] Tang J, Zhang J, Yao L, et al. ArnetMiner: Extraction and mining of academic social networks [C] //Proc of the 14th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2008: 990-998

[3] China Computer Federation. CCF Scholar [EB/OL]. [2015-12-02]. http://www.ccf.org.cn/sites/ccf/ccfscholar.jsp (in Chinese)
(计算机学会. CCF学术空间[EB/OL]. [2015-12-02]. http://www.ccf.org.cn/sites/ccf/ccfscholar.jsp)

[4] Chen Wei, Wang Zhongyuan, Yang Sen, et al. ScholarSpace: An academic space for computer science researchers [J]. Journal of Computer Research and Development, 2011, 48(S3): 395-399 (in Chinese)
(陈威, 王仲远, 杨森, 等. ScholarSpace:面向计算机领域的学术空间[J]. 计算机研究与发展, 2011, 48(S3): 395-399)

[5] Erickson J. On the relative complexities of some geometric problems [C/OL] //Proc of Canad Conf Computer Geometry. 2010: 85-90. [2015-12-01]. http://cs.uiuc.edu/~jeffe/pubs/pdf/relative.pdf

[6] Erickson J. New lower bounds for Hopcroft's problem [J]. Discrete & Computational Geometry, 1996, 16(4): 389-418

[7] Gan Junhao, Tao Yufei. DBSCAN revisited: Mis-Claim, un-Fixability, and approximation [C] //Proc of the 2015 ACM SIGMOD Int Conf on Management of Data. New York: ACM, 2015: 519-530

[8] Shi L, Tong H, Tang J, et al. VEGAS: Visual influence graph summarization on citation networks [J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27(12): 1-15

[9] Ao Wenjing. Study of mining social network based on call records [D]. Guangzhou: South China University of Technology, 2013 (in Chinese)
(敖文井. 基于通话记录的社会关系网络挖掘[D]. 广州: 华南理工大学, 2013)

[10] Zhang J, Fang Z, Chen W, et al. Diffusion of "Following" links in microblogging networks [J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27(8): 2093-2106

[11] Tang J, Wang B, Yang Y, et al. PatentMiner: Topic-driven patent analysis and mining [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012: 1366-1375

[12] Sun J, Tang J. A Survey of Models and Algorithms for Social Influence Analysis [M] //Social Network Data Analytics. Berlin: Springer, 2011: 177-214

[13] Kumar R, Mahdian M, McGlohon M. Dynamics of conversations [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 553-562

[14] Gruhl D, Guha R, Liben-Nowell D, et al. Information diffusion through blogspace [C] //Proc of the 13th Int Conf on World Wide Web. New York: ACM, 2004: 491-501

[15] Mishra N, Schreiber R, Stanton I, et al. Finding strongly knit clusters in social networks [J]. Internet Mathematics, 2008, 5(1): 155-174

[16] Liben-Nowell D, Kleinberg J. Tracing information flow on a global scale using Internet chain-letter data [J]. Proceedings of the National Academy of Sciences of the United States of America, 2008, 105(12): 4633-4638

[17] Yang Z, Guo J, Cai K, et al. Understanding retweeting behaviors in social networks [C] //Proc of ACM Conf on Information & Knowledge Management. New York: ACM, 2010:1633-1636

[18] Myers S A, Zhu C, Leskovec J. Information diffusion and external influence in networks [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012:33-41

[19] Matsubara Y, Sakurai Y, Prakash B A, et al. Rise and fall patterns of information diffusion: Model and implications [C] //Proc of the 18th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2012:6-14

[20] Huang H, Tang J, Liu L, et al. Triadic closure pattern analysis and prediction in social networks [J]. IEEE Trans on Knowledge & Data Engineering, 2015, 27(12): 3374-3389

[21] Tang J, Lou T, Kleinberg J. Inferring social ties across heterogeneous networks [C] //Proc of the 6th ACM Int Conf on Web Search & Data Mining. New York: ACM, 2012: 743-752

[22] Wang C, Han J, Jia Y, et al. Mining advisor-advisee relationships from research publication networks [C] //Proc of the 16th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2010: 203-212

[23] Backstrom L, Leskovec J. Supervised random walks: Predicting and recommending links in social networks [C] // Proc of the 4th ACM Int Conf on Web Search & Data Mining. New York: ACM, 2010: 635-644

[24] Diehl C P, Namata G, Getoor L. Relationship identification for social network discovery [C] //Proc of the 22nd National Conf on Artificial Intelligence. Menlo Park, CA: AAAI, 2007: 546-552

[25] Eagle N, Lazer D. Inferring Social Network Structure Using Mobile Phone Data [M]. Berlin: Springer, 2008

[26] Sun Y, Tang J, Han J, et al. Co-evolution of multi-typed objects in dynamic star networks [J]. IEEE Trans on Knowledge & Data Engineering, 2014, 26(12):2942-2955

[27] Zhang J, Liu B, Tang J, et al. Social influence locality for modeling retweeting behaviors [C] //Proc of the 23rd Int Joint Conf on Artificial Intelligence. New York: ACM, 2013: 2761-2767

[28] Burt R S. Structural Holes: The Social Structure of Competition [M]. Cambridge, MA: Harvard University Press, 1992

[29] Goyal S, Vega-Redondo F. Structural holes in social networks [J]. Journal of Economic Theory, 2007, 137(1): 460-492

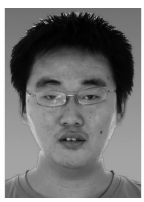
[30] Lou T, Tang J. Mining structural hole spanners through information diffusion in social networks [C] //Proc of the 22nd Int Conf on World Wide Web. New York: ACM, 2013: 825-836

[31] Ester M, Kriegel H, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C] //Proc of the 3rd ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 1996: 226-231

[32] Gunawan A. A faster algorithm for DBSCAN [D]. Eindhoven, the Netherlands: Technische University Eindhoven, 2013

[33] Han J, Kamber M, Pei J. Data Mining: Concepts and Techniques [M]. San Francisco, CA: Morgan Kaufmann, 2012

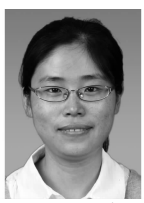
[34] Hjaltason G R, Samet H. Index-driven similarity search in metric spaces [J]. ACM Trans on Database Systems, 2003, 28(4): 517-580



He Xianmang, born in 1981. Received his PhD degree from the School of Computer Science, Fudan University. Lecturer in the Faculty of Information Science and Engineering of Ningbo University. His main research interests include database, coding and cryptography.

Chen Yindong, born in 1983. Received his PhD degree from the School of Computer Science, Fudan University. Associate professor in the College of Engineering, Shantou University. His main research interests include information security and cryptology.

Li Dong, born in 1970. Received her master degree from the Department of Computer Science and Technology, Tsinghua University. Senior engineer in National Natural Science Foundation of China. Her main research interests include database technology and information system management.



Hao Yanni, born in 1978. Received her master degree from the School of Computer Science and Engineering, Beihang University. Engineer in National Natural Science Foundation of China. Her main research interests include database technology and information system management.