“智慧校园”学者画像系统研究

彭程程 吴斌
（北京邮电大学智能通信软件与多媒体北京市重点实验室，北京 100876）

摘要：随着大数据时代的到来，学术相关数据呈指数增长趋势。同时，用来刻画用户行为的用户画像，近期在各个领域得到了广泛应用。通过分析挖掘与学者相关的学术数据，可以对学者进行全方位、高精度的画像构建，这对研究学者的学术行为有重要的作用。本文介绍了“智慧校园”学者画像系统及系统的相关技术点与功能特色，并将其与其他主流学者画像系统进行对比分析。结果显示，该系统在研究学者的学术谱系、研究脉络等方面存在一定的优势与特色。

关键词：学者画像；用户画像；学术谱系；六度搜索
中图分类号：TP182
DOI：10.3772/j.issn.1673-2286.2019.02.001

随着科学技术的迅速发展，学术相关数据呈现了指数增长趋势，现代社会对科研人才的需求也越来越明显。虽然用户可以在网络上获取到科研人员相关信息，但是海量的网络学术信息使得科研人才信息分布零散，用户不能直观地获取学者的相关信息。因此，快速、精确、全面地获得学者信息成了亟待解决的问题。

用户画像是对真实用户的抽象描述方式，通过构建多维度标签属性来描述用户或用户群的兴趣、特征、行为及偏好，从而为产品优化、精准营销、个性化服务等提供数据支撑。近年来，国内外学者在用户画像领域做了大量的调研工作，并取得了一定的研究成果，这在学术界与产业界都具有重大意义。但是很少有研究人员能针对学者的个人特征及学术行为特征进行深入、精细的描述和刻画。

学者画像系统可以相对明确地展示学者的基本信息、研究方向、社交关系，甚至整个行业的研究趋势，这对于互联网时代的科研发展和专家遴选非常重要。以研究学者为中心的学术智库已在国家自然科学基金委员会、科技部、中国工程院等权威机构展开了应用[1]。

传统的学者画像从学者论文发表情况的角度出发，只粗糙地对学者进行表层刻画，如过于简略的个人信息、粗浅的学者合著网络、不够丰富直观的学术关键词、无法描述学者的学术谱系等。因此，本系统对多源数据进行分析，使用实体消歧、数据融合等文本分析方法和社团发现等数据挖掘方法，对学者和机构进行建模，多维度挖掘学者的深层学术信息。系统通过展示学者详细的个人信息、丰富的合作关系、传承的学术谱系、六度搜索路径、关键人物的发现与替代等功能，刻画更真实、更准确、更立体的科研学者，为专家遴选、学术热点分析等提供数据支持。

1 相关工作

1.1 学者画像系统

目前，传统学者画像主要以各机构产出系统的形式进行呈现。谷歌学术、百度学术、万方数据知识服务平台、中国知网、dblp、Aminer、c-dblp、科搜、Web of Science、Engineering Village、ACM Digital Library等平台均对学者进行了画像构建。

谷歌学术是影响力最大的学术搜索网站[2]。Web of Science拥有全球最大、覆盖学科最广的学术资源。Engineering Village是全球最权威的工程与应用科学领域的文献检索平台。这3个平台侧重于论文检索，学者画像功能比较单一，主要功能点包括简略的个人信息与论文发表情况，缺乏对学者更丰富的刻画。ACM

Digital Library集合了ACM和5 000多家出版社的出版物，旨在为专业和非专业人士提供了解计算机和信息技术领域资源的窗口，其学者画像功能包括学者的简略个人信息、发表论文、研究关键词、相关学者等，其学术评价指标维度较为丰富新颖。百度学术是一个提供中英文文献检索的学术资源搜索平台，涵盖各类学术期刊、会议论文[3]。万方数据知识服务平台整合国内外学术资源，集成期刊、学位论文、会议论文、科技报告、专利、视频等十余种资源类型。中国知网提供中国学术文献、外文文献、学位论文、报纸、会议论文、年鉴、工具书等各类资源统一检索、统一导航、在线阅读和下载服务。这3个平台同样侧重于论文检索，学者画像功能较为单一，功能主要包括学者的学术评价指标、发表论文、合作学者、合作机构。dblp与AMiner平台针对计算机科学领域，为用户提供该领域学者的相关信息。dblp是德国特里尔大学搭建的计算机科学文献检索网站，其中涵盖计算机学术会议、期刊、报告、书籍在内的海量文献记录，便于科研人员查询计算机领域相关文献信息，其权威性得到了研究界的高度认可。但是，dblp没有提供对中文文献的收录和检索功能，其学者画像功能包括学者发表论文、合作学者。AMiner是研究者学术搜索类网站，为计算机科学相关领域的研究者提供领域知识[4]，主要功能模块有个人信息、研究兴趣、合作学者、发表论文、学术评价指标、学者迁徙路线。该系统功能更为丰富，但同样也存在学者个人信息过于简略、不能描述学者的学术谱系等不足。c-dblp是由中国人民大学开发的基于中文论文的学术信息集成系统，包括ScholarSearch、ScholarTree、ScholarExplorer、ScholarGraph和ScholarRankings5个子系统，其中ScholarExplorer子系统是以作者为中心的学者画像系统，主要功能模块包括个人信息、研究兴趣、合作学者、发表中文论文、师承关系等。科搜是国家科技资源共享服务工程技术研究中心支持的学术搜索网站，主要功能模块包括个人信息、研究兴趣、学术圈、相关论文、相关获奖等。

综上所述，在用户画像的建模过程中，研究者对于立体精准的学者画像构建研究较少。立体是指描述用户的标签维度多，精准是指描述用户的标签准确，能够准确地描述科研人员的各种特性，通过构建立体精准画像保证从多个角度接近最真实的用户。目前的画像构建方法已不能很好地解决这些问题。

与上述主流学者画像系统比较，“智慧校园”学者画像系统具有8个特点：①学者个人信息属性维度较多；②可从时间维度出发，展示学者的学术关键词变化趋势；③相关学者信息较为丰富、直观；④具有机构社团发现功能；⑤具有机构关键词变化趋势功能；⑥可以描述多层的学术谱系，脉络较为清晰；⑦具有六度搜索路径功能；⑧具有团队核心人物演化分析功能。

1.2 相关技术

1.2.1 信息抽取

学者画像系统需要从结构化数据与非结构化数据中抽取出学者的个人信息、教育经历、所在机构、联系方式等。当前主流的方法主要包括基于序列标注的方法和基于关系抽取模型的方法。

基于序列标注的方法大多依赖条件概率模型。信息抽取常用模型包括最大熵Markov模型、条件随机场模型、动态条件随机场模型、树状条件随机场模型等。

基于关系抽取模型的方法将学者信息抽取问题转化为关系抽取问题。近年来，深度学习被广泛应用于关系抽取领域，Zhou等[5]将LSTM与词级别的注意力机制相结合，Lin等[6]将CNN与句子级别的注意力机制相结合，两者的模型均取得了较大的提升；Yang等[7]将多个LSTM分类器组合在一起，进一步提高模型效果。

学者信息抽取是构建学者画像的基础工作。随着互联网数据指数级增长，信息抽取技术也逐渐从面向特定领域、特定信息的基于人工模板的方法转变为面向开放领域的开放式信息抽取方法。

1.2.2 重名消歧

学者画像系统构建中的同名消歧问题一直被视为一个具有挑战性的问题，学术文献数量的飞速增长使得该问题变得更加困难与紧迫。尽管同名消歧已经在学术界和工业界被大量研究，但该问题仍未能很好地解决。姓名消歧问题主要通过基于特征抽取的聚类与基于链接的聚类两种方法进行解决。

基于特征抽取的聚类方法通常采用有监督的方法在文档之间根据其特征向量学习一个正确的距离函数。Yoshida等[8]提出了两阶段聚类算法，第一阶段是采用凝聚聚类方法的强聚类，第二阶段分别采用强聚类和弱聚类提高聚类结果的召回率。Louppe等[9]使用了一个分类器来学习两实体之间的相似度，这种方法取得

了比半监督层次化聚类更好的效果。Zhang等[10]提出综合全局监督和局部内容的网络表示学习框架及端到端的聚类大小估计算法来获取更好的消歧结果。

基于实体链接的方法可以利用图的拓扑性质和来自邻居节点的聚合信息进行消歧。Fan等[11]使用合作者信息作为输入，通过对作者合作关系图的构建，进而进行有效路径选择及相似度计算，最终完成聚类。Tang等[12]在统一概率图框架中，采用隐马尔可夫随机场对节点和边的特征进行构建。Zhang等[13]尝试基于文档相似度和合作关系，通过对3个已经构建好的网络进行表示学习。

重名消歧工作是学者画像系统中科研人物搜索、学者兴趣挖掘、科学文献管理、社交网络分析等方面的基础工作。

1.2.3 社区发现

学者画像系统中的社区发现问题可以从网络的拓扑结构中发现潜在的学术群体化结构特性，有助于观察和研究整个学者关系网络。

在静态社区发现方面，Yin等[14]尝试通过合并网络结构获取高阶的网络信息表示来处理有向网络；Epasto等[15]提出一种自我网络分裂框架，通过非重叠算法实现重叠社区发现。在动态社区发现方面，Folino等[16]提出了基于遗传表示的算法来平衡最大化聚类精度与最小化两个相邻时间片之间的聚类差异；Ma等[17]通过构建进化非负矩阵分解框架在不增加时间复杂度的情况下，寻找全局最优解，避开局部最优解；Niu等[18]将标签传播思想引入到动态社区检测多目标优化算法中，提高社区发现质量与收敛速度。在大规模并行社区发现方面，Wu等[19]提出了一种基于距离动态的大规模并行社区检测算法PCDU，该方法适用于大规模网络中的社区划分结果评价；Zhang等[20]提出一种基于增量计算的并行动态非重叠社区发现算法PICD，充分利用网络短时平滑性特点，通过不断优化网络的PWCC来获取高质量的社区结构。

2 “智慧校园”学者画像系统构建

“智慧校园”不仅提供了学者检索、学者发表论文、学者合作关系等学者画像系统基本功能，还通过抽取和分析机构官网、学者发表论文数据、学位论文数据等多源数据，深入挖掘学者的详细个人信息、研究领域、学术关键词、学术谱系、六度搜索路径等信息，为科研评价和决策提供更多可信赖的依据。

2.1 系统架构

2.1.1 系统技术架构

“智慧校园”学者画像系统设计模式采用MVC模式，其耦合性较低、可重用性较高、部署速度快、可维护性较高。前端开发使用HTML、CSS、JavaScript 3种语言，应用Bootstrap和jQuery两个前端开发框架。Bootstrap框架可以提高开发效率、便于后期维护、规范项目开发流程，同时也可以使CSS代码更加简明易懂，让HTML代码更规范合理。系统后台开发使用SpringMVC框架，可以让开发流程变得层次清晰。系统后台开发使用Java语言，具有可解释、可移植、多线程、动态性等优点[21]。数据存储使用Neo4j，Neo4j是一个NoSql数据库，用于网络图的存储，它对数据库的操作更迅速，数据显示方式更加直观、灵活。

2.1.2 系统层次架构

“智慧校园”学者画像系统架构层次从低层到高层共分为三层，即数据支撑层、文本挖掘层、数据可视化层。

数据支撑层是系统架构的最底层，包括数据的采集和存储。数据源分为开源数据和闭源数据两种；数据采用Neo4j数据库存储。文本挖掘层用来完成系统中重要的数据处理任务，包括实体识别与融合、关系发现、关键词抽取、社团发现等，对学者和机构进行建模。数据可视化层是系统与用户交互的核心，以功能模块的方式展示学者的个人信息、发表论文、研究关键词、研究趋势、合作关系、学术谱系、六度搜索路径，以及关键学者的发现与替代及机构的关键词、研究趋势、社团划分信息。具体架构如图1所示。

2.2 网络融合

在传统的社会科学领域，社会关系的多重性被用来表征用户之间社会交换关系的多个方面。关系多重性的思想可以推广到各种网络中。在数据挖掘领域，使用

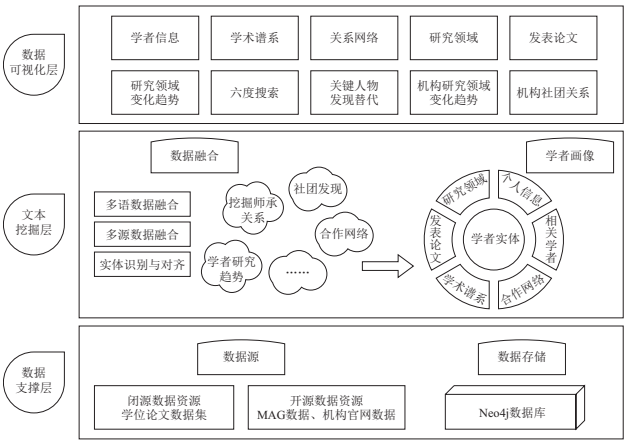

图1 系统层次架构设计

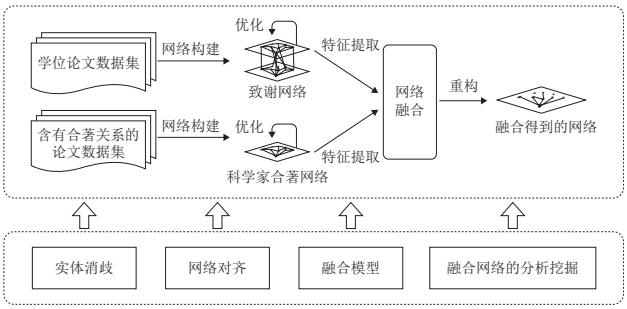

图2 网络融合流程

“多关系网络”来表示社交网络中的多类型关系，有助于数据挖掘任务的进行。

关系网络是知识服务平台的必要内容。传统的学者关系网络即为“科研合著网络”，主要采用论文署名中学者“共现”方法对学者之间的关系进行建模[22]。然而，这种方法是粗糙的，学者之间的关系不仅包括论文合著关系，还包括项目合作关系、共同指导学生等关系。因此，在传统的学者画像系统中，作者之间深层的关系没有得到精确的刻画。

因此，我们用学位论文数据集中的致谢数据对传统的学者关系网络进行深层次刻画。不同学者在同一篇学位论文致谢部分的共现很大程度上体现了学者之间的工作合作关系，如共同指导学生关系、项目合作关系、共事关系等。我们将传统的、粗糙的共现合著关系网络结合相应的领域深层数据进行融合分析，以此构建出更真实、更准确的学者关系网络。

如图2所示，首先，进行命名实体抽取工作。我们使用多语言实体之间的映射和命名实体消歧等技术，从多个数据源中识别学者实体和连边之间的关系作为图中的节点之间的连边，从而为多层网络的构建提供数据支撑。接着，进行多层网络的构建。为了保存多网络结构，需要对多网络进行数据结构的存储和网络之间关系的建模。最常使用的是多层网络，该类型网络不仅能够保存多网络的结构特征，还能够对网络之间的相互依赖进行建模。然后，我们对致谢数据集的学位论文致谢部分抽取学者合作网络；对MAG（Microsoft Academic Graph）论文数据集抽取学者合著网络。最后，进行网络融合。在构建的致谢网络中以标注的社区结构为标准数据集，利用随机梯度下降算法（GBDT）实现半监督的网络融合。

2.3 师承关系挖掘

学术谱系是由学术传承关系（包括师承关系）关联在一起的、不同代际的科学家所组成的学术群体[23]。对学术谱系的挖掘，旨在构建并深入挖掘各门学科或主要学科分支层面上学术谱系的产生、运作、发展的过程及一般趋势，促进一流学术谱系的传承以及科研人才的培养[24]。学者学术谱系为分析学者之间的互动提供了至关重要的信息，也可以为研究者提供许多具体的应用，如学术顾问推荐、学术新星挖掘等[25]。

学位论文中蕴含着丰富而准确的学术谱系关系。通过收集大量的学位论文并应用实体抽取、关系抽取等技术，可以挖掘出时间跨度大、覆盖范围广、准确度高的学术谱系。如图3所示，“智慧校园”系统首先通过对含有论文指导关系的结构化数据进行<导师，指导关系，学生>三元组的构建，接着使用基于自定义词典的HanLP中文自然语言处理工具包对缺失结构化论文指导关系的文本数据进行实体识别、关系抽取以获得三元组。对三元组集合采用基于逻辑规则的关系推断方法识别潜在的师承关系，进一步整理得到最终的多层学术谱系。

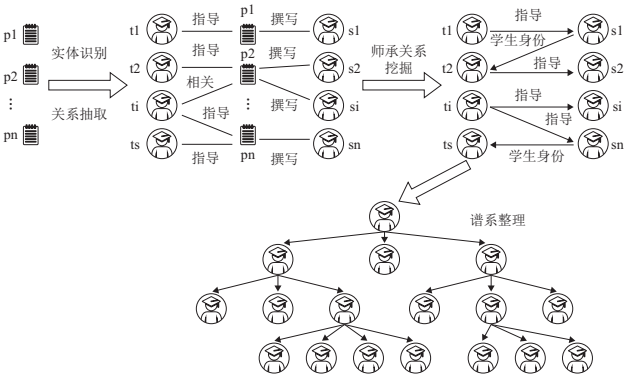

图3 师承关系挖掘流程

2.4 团队核心人物演化分析

团队核心人物演化分析在学者画像方面的应用在于：预测学术机构内某位学者离职后，哪位学者会接替他的位置。使用学院内学者的职位变更记录来模拟网络的演变，以学者的科研水平模拟网络的影响力。作出如下假设：①科研能力越高对应网络的影响力越大，越有可能成为替代者；②科研能力通过教师发表的科技论文和网络内部合作体现；③学者之间的关系是通过论文合作关系体现的；④学者的级别是根据教师的职称来确定的；⑤职位变动的替代者来自网络内部；⑥职位变动依据的是学者在网络内部的科研能力。

算法主要分为3个部分。①继任者问题（TSP）。当某个学者r离职后，算法将推算r被另一名学者v替代的可能性。②网络重塑问题（TNRP）。根据每个点的影响力计算网络的整体影响力，并且确定出一个需要从网络中清除的学者集合k，以此最大程度地降低剩下网络的预期运作效力。③多学者继任者问题（MTSP）。当从一个网络中移除了多名学者后，算法将推算可能会诞生的新网络以及相关的概率分布。算法流程图如图4所示。

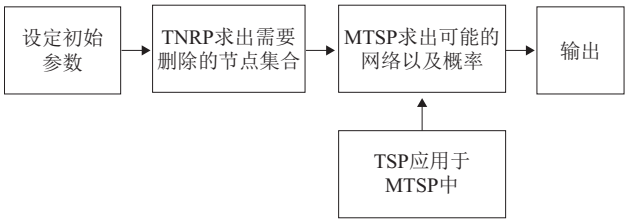

图4 关键人物发现与替代算法流程

在算法流程中，TNRP为了决定要从网络中移除的k个顶点，首先定义可能的网络。求解TNRP的过程十分复杂，我们采用了一种贪心算法以进行求解，其输入为网络及要移除的学者个数k，输出为要删除的顶点集合。如图5所示的TNRP算法流程图所示，每一次循环遍历图中的每一个节点，假设要删除该节点并计算删除之后的网络影响力，得到使网络影响力最小的节点，并真正删除该节点。循环k次得到需要删除的节点集合。

MTSP算法流程图如图6所示。要移除的顶点集合的候补者替代者集合中，必须满足4个条件：①每个被删除顶点具有一个替代顶点；②替代顶点无法被移除；③替代节点必须为候补者；④候补者只能替代一个顶点。

MTSP在同时寻找多名重要性高的学者情况中，可能会存在许多网络（如果有n个人可以替代a，m个人可以替代b，那么就可能会出现nm个新网络）。MTSP的设计目标是帮助确定新网络的概率。

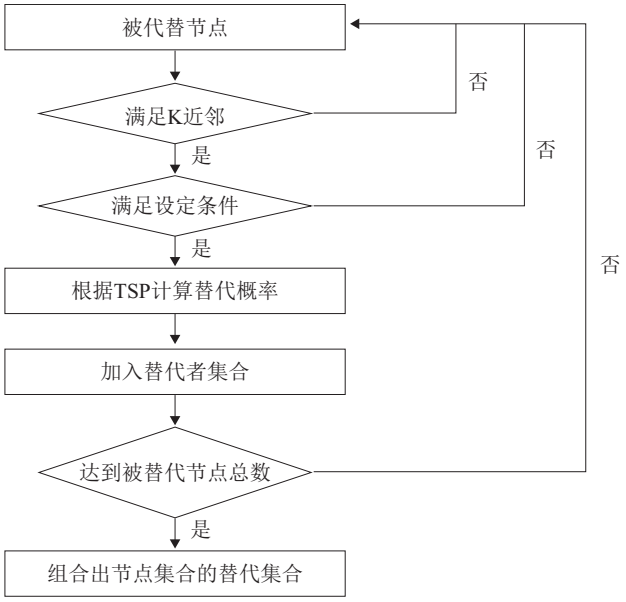

图5 TNRP算法流程

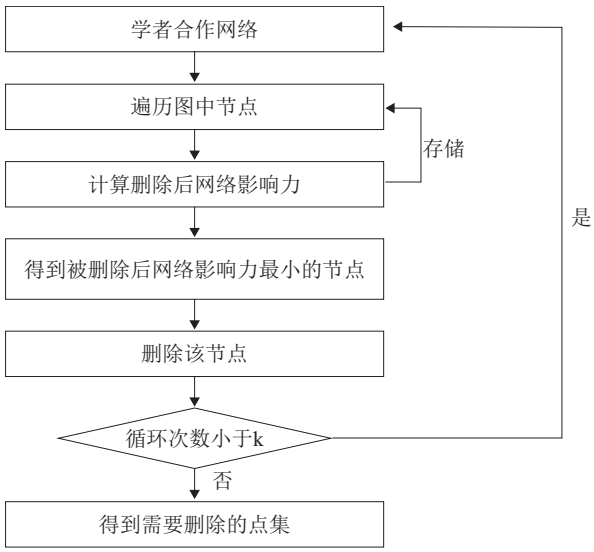

图6 MTSP算法流程

2.5 机构社团划分

机构社团划分可以从社区发现的角度对机构成员之间的关系作出直观的刻画，以此得到学者所在的学术群体。我们通过使用Fast Unfolding算法进行机构社团发现。Fast Unfolding算法的流程：①先将图中每个节点都看作一个独立的社团，初始社团的数目即为节点的数目；②对初始社团中的每个节点i，依次尝试把该节点分配到其每个邻居节点所在的社团，计算分配前后的模块度变化ΔQ，并记录模块度变化ΔQ最大的那个邻居

节点，如果max（ΔQ）>0，则把该节点分配到ΔQ最大的那个邻居节点所在的社团，否则该节点所属社团保持不变；③重复步骤②，直到所有节点所在的社团不再进行变化；④对图根据社团进行压缩，将所有在同一个社团的节点压缩成一个新的节点，社团内节点之间的边权重转化为新产生的节点的环权重，社团间的边权重转化为新节点之间的边权重；⑤重复步骤①直到整个图的模块度不再发生变化[26]。

3 “智慧校园”学者画像系统实例分析

3.1 数据集

系统依托的数据包括闭源数据和开源数据两种类型。闭源数据为北京邮电大学高校硕士生及博士生毕业论文数据集中的致谢部分；开源数据包含两个数据集，MAG数据集和机构官网数据集。①高校硕士生及博士生毕业论文致谢数据集是1997—2015年硕士、博士学位论文致谢章节的集合。其中每个实体为一篇毕业论文，从中可以获取到论文题目、作者姓名、作者所在高校、作者所在专业、指导老师姓名、论文关键词、论文致谢部分内容，其中包含了大量的人物信息及人物实体之间的关系，其语言为中文。②MAG数据集是微软学术提供关于论文的数据集[27]，其中每个实体为一篇论文，我们从中可以获取到论文题目、作者姓名、作者所在单位、论文发表年份、关键词、研究领域信息。③机构官网数据包括机构官网中对学者个人信息的描述和涉及到机构新闻公告信息。

3.2 系统功能模块

3.2.1 个人信息模块

该模块功能点为展示学者的个人信息，包括学者的姓名、性别、照片、供职机构、所在中心、职称、职务等信息。“智慧校园”系统利用爬虫及文本分析方法，在北京邮电大学各学院官网抓取到信息与通信工程学院、电子工程学院、计算机学院、自动化学院、软件学院、数字媒体与设计艺术学院、现代邮政学院、网络空间安全学院、光电信息学院、理学院、公共管理学院、人文学院、马克思主义学院、国际学院这14个学院中导师的个人网页及相关校内新闻。

传统的学者画像系统获取学者个人信息的方式为：①从学者发表论文中提取学者的姓名、联系方式、供职机构等信息；②从互联网中抽取学者相关的个人信息。这些方式存在着一些不足，如第一种方式获取的学者信息量过少，无法对学者进行多维刻画；第二种方式获取的学者信息量较多，同一属性可能存在多个属性值，对学者进行精准刻画的概率较低。

从官网中获取的学者信息拥有比从论文中抽取的信息更加丰富的维度，如性别、职称、职务等。同时，从官网中获取的学者信息也会更新较快、更准确。本系统通过解析机构官网上的学者信息，进行学者个人信息属性的挖掘，从而得到更精准、更多维的学者个人信息。

3.2.2 学者关系网络

在传统的学者画像系统中，主要采用论文署名中学者“共现”方法对学者之间的关系进行建模，构建出的网络为“学者合著网络”，不能对学者之间的项目合作关系、共同指导学生等关系进行更精细描述。

因此，我们对学位论文数据集中的致谢数据与学者发表论文数据进行联合挖掘。对致谢数据集的学位论文致谢部分抽取学者合作网络；对MAG论文数据集抽取学者合著网络。最后进行网络融合，在构建的致谢网络中以标注的社区结构为标准数据集，利用随机梯度下降算法（GBDT）实现半监督的网络融合。

3.2.3 学术关键词模块

学者的研究领域及学术关键词是学者画像中重要的组成部分，能够充分体现出学者的研究方向、学术偏好，甚至可以体现出学者对该学科领域热点的关注度。传统的学者画像系统，如谷歌学术、百度学术等，不涉及学术关键词功能；少数学者画像系统，如AMiner、中国知网等，只是对学术关键词进行简单的罗列。这种描述方式并不能描绘学者在不同时间段关注的学术关键词及学者的研究路线变迁趋势。

因此，我们从时间维度出发，对学者的学术关键词进行刻画，将学者研究关键词的变化趋势直观地体现出来，从而发现学者的学术研究脉络。学术关键词变化趋势如图7所示。

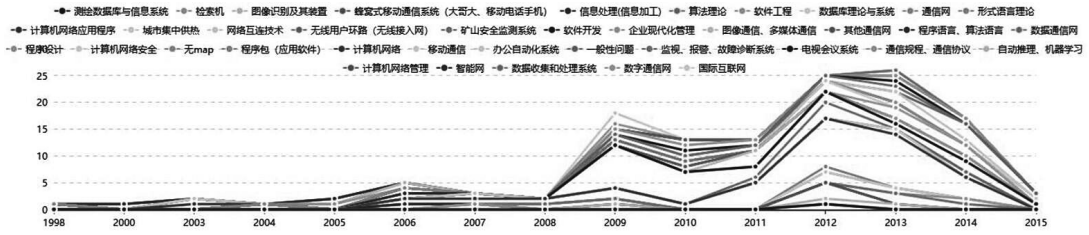

图7 学术关键词变化趋势

3.2.4 学术谱系

传统的学者画像系统很少对学者的学术谱系进行描述。“智慧校园”系统以学位论文为数据集，使用实体识别与关系抽取等方法从中挖掘实体及其链接关系，使用关系推断等方法从网络中识别出潜在的师承关系，进一步归纳整理得到最终的学术谱系。如图8所示，这是北京邮电大学陈俊亮院士的多层学术谱系，从中可以看出，陈俊亮指导的学生王柏作为导师指导了另一批学生。

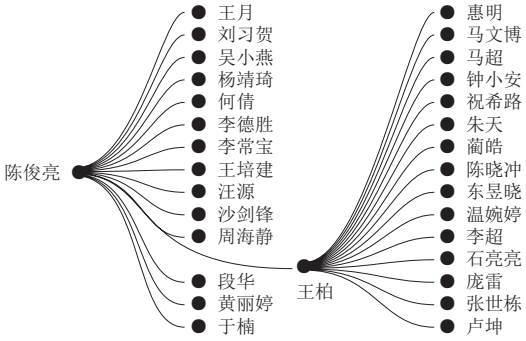

图8 学术谱系模块

3.2.5 六度搜索

六度搜索的含义是指现实生活中的任意两个人之间建立起关联最多只需要通过六个人。在“六度分离”理论中，社会中普遍存在人与人之间的弱纽带关系，这种关系能够拉近互不相识的两个人之间的距离，这在社会关系中发挥着巨大的作用。“六度分隔”产生的关系路径可以利用熟人之间的联系产生一个可信任的网络，这其中的潜能的确是无可估量的。

传统的学者画像系统不包含六度搜索功能。我们根据用户输入的两个人物实体对象，发掘实体间的关联路径及其路径中的人物。首先根据需要查询的人物关系从关系网络中进行实体搜索，再通过图算法获取极大连通子图作为网络关系的查询结果。本系统的六度搜索功能产生一个可信任的网络路径，通过这个网络路径，能够更清晰明确地观察到两名学者在科研关系网络中的信任路径，为他们提供潜在的合作可能。六度搜索网络如图9所示。

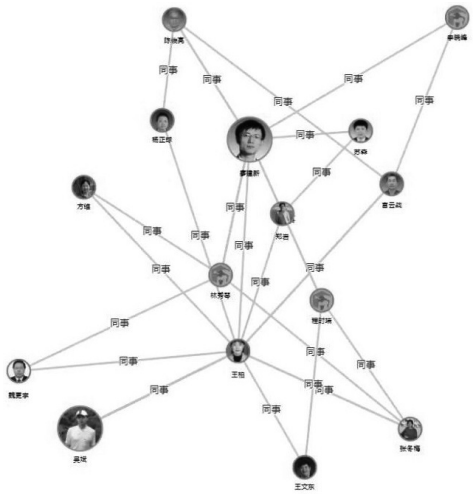

图9 六度搜索网络

3.2.6 团队核心人物演化分析模块

传统的学者画像系统缺少团队核心人物演化分析功能。团队核心人物的挖掘预测是对科研团队群体行为的分析，“智慧校园”学者画像系统够预测学术机构内某位学者离职后，哪位学者会接替他的位置。我们通过解决TNRP网络重塑问题求出网络中需要删除的学者集合，接着解决MTSP多学者网络继任者问题来预测出可能构建的新网络及相关的概率。当科研合作网络中移除一名团队核心人物时，该模块可以预测出网络的演化结果以及核心人物的更替。

3.2.7 机构研究趋势

机构研究趋势是某所机构研究重心、研究热点的直观体现。机构的研究趋势也会间接影响学者未来的研究方向。传统的学者画像系统，如谷歌学术、百度学术、中国知网、AMiner等都不包含对机构学术研究关键词研究趋势的挖掘。同样从时间维度出发，我们由机

构名下学者的学术关键词趋势上卷得到该机构的学术热词的演化趋势。

3.2.8 机构社团划分

传统的学者画像不包含机构社团划分功能，我们通过使用Fast Unfolding算法进行机构社团发现，将刻画两名学者关系的“一对一层次”拓展到刻画多名学者关系的“多对多层次”。这样，可以更深入、更直观地挖掘机构名下的科研团队信息。

4 系统对比

将本系统与其他学者画像系统进行功能对比，结果见表1。“智慧校园”学者画像系统存在以下优势与特点：个人信息属性维度较多；能够更直观地展示在时间维度上的关键词变化趋势；相关学者信息较为丰富；有机构社团发现功能；有机构关键词变化趋势功能；学术谱系脉络较为清晰；有六度搜索路径功能；有关键人物发现与替代功能。同时本系统也存在着一些缺点，如数据量较小、没有对论文引用关系进行描述等。

5 总结

本文首先讨论了学者画像在互联网时代的重要意义，接着介绍了传统学者画像的功能特点及其存在的局限性，重点讲述了“智慧校园”学者画像系统的系统架构、数据集和功能模块的实现与可视化。本系统对多

表1 “智慧校园”与其他学者画像系统对比

	个人信息	评价指标	论文信息	关键词趋势	相关学者	机构关键词研究趋势	机构社团发现	其他功能点
“智慧校园”系统	姓名、照片、性别、所在机构、所在中心、职称、职务、学者简介、研究领域信息	发表论文数，合作学者数	发表论文信息	关键词丰富，折线图展示随年份变化趋势	以关系图形式展示，数目较多	较为丰富，以折线图形式展示随年份变化趋势	有社团发现功能，且区分出的社团和真实分组情况较为接近	多层的学术谱系功能、六度搜索功能、关键人物发现与替代功能
谷歌学术	姓名、所在机构、所在中心、研究领域信息	被引频次、发表论文数、i10指数	发表论文信息。论文数目较多	无	以列表形式展示，数目较少	无	无社团发现功能，以相关作者方式展现	无
百度学术	所在机构、所在中心、研究领域信息	被引频次、成果数、h指数、g指数	发表论文信息。论文数目较多	无	以关系图形式展示，数目较多	无	无社团发现功能，以相关作者方式展现	合作机构功能；按期刊、会议、专著统计成果
中国知网	姓名、所在机构、研究领域信息	发表论文数、总下载量	发表论文信息。论文数目较多	无	以列表和关系图形式展示，数目较多	无	无	有学者的导师、学生列表，无师承树；学者所获基金功能
万方数据知识服务平台	姓名、所在机构、研究领域信息	被引频次、发表论文数、h指数、关注者数	发表论文信息	无	以关系图形式展示、数目较多	无	无社团发现功能，以相关作者方式展现	相关学者功能；按年份进行个人成果数、被引数统计
dblp	姓名、所在机构/中心、教育经历	发表论文数	发表论文信息	无	以列表形式展示，数目较少	无	无	无
c-dblp	姓名、所在机构、研究领域信息	被引次数	发表论文信息（中文）	无	以关系图形式展示	无	无	承担项目功能、学术谱系功能
科搜	姓名、所在单位、职称、研究领域、教育经历、成就、所获奖项、社会荣誉信息	发表论文数、相关项目数	发表论文信息	无	百科关系、中文论文合作关系、英文论文合作关系、专利合作关系、同行人物	无	无	相关专利功能、相关获奖功能

续表

	个人信息	评价指标	论文信息	关键词趋势	相关学者	机构关键词研究趋势	机构社团发现	其他功能点
AMiner	姓名、照片、职称、所在机构、研究中心、教育经历、研究领域信息	被引频次、发表论文数、合作学者数、最近活跃度、发文丰富度、h指数、g指数	发表论文信息	关键词较少，折线图展示随年份变化趋势	以关系图形式展示，同一个学者出现多次	无	无社团发现功能，以相关作者方式展现	学者迁徙路线功能、有学者专利功能、相似作者功能、增加D-core评价指标
Web of Science	姓名、所在机构	出版物总数、h指数、每项平均引用次数、被引频次总计	发表论文信息。论文数目较多	无	无	无	无	基金资助机构功能、作者的论文引用报告功能
Engineering Village	姓名、所在机构、研究领域信息	无	发表论文信息。论文数目较多	无	无	无	无	作者的论文引用报告功能
ACM Digital Library	姓名、所在机构、教育经历、研究领域信息	论文被引数、每项平均引用次数、出版物总数、出版年份区间、可下载论文数、每篇文章平均下载量等	发表论文信息。论文数目较多	无	以列表形式展示	无	无社团发现功能，以相关作者方式展现	抽取作者官网主页内容，作为一个功能模块

源数据进行分析，使用实体消歧、数据融合等文本分析方法和社团发现等数据挖掘方法，对学者和机构进行建模，从多维度挖掘学者的深层学术信息。本系统通过展示学者多属性的个人信息、丰富的合作关系、传承的学术谱系、六度搜索路径等，为用户刻画更真实、更准确、更生动的科研学者。最后，本文将“智慧校园”学者画像系统与主流的其他学者画像系统进行功能对比，直观体现出了本系统的特色功能。

参考文献

[1] 袁莎，唐杰，顾晓韬. 开放互联网中的学者画像技术综述 [J]. 计算机研究与发展，2018，55（9）：1903-1919.

[2] ANNEWIL H，SATU A. Google Scholar，Scopus and the Web of Science：A Longitudinal and Cross-Disciplinary Comparison [M]. New York：Springer-Verlag New York，2016.

[3] 魏瑞斌，郭一娴. 基于用户体验的百度学术应用研究 [J]. 现代情报，2017（5）：91-99.

[4] TANG J，ZHANG J，YAO L，et al. Arnetminer：Extraction and Mining of Academic Social Networks [C] //Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas：ACM，2008.

[5] ZHOU P，SHI W，TIAN J，et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification [C] //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin：ACM，2016.

[6] LIN Y，SHEN S，LIU Z，et al. Neural relation extraction with selective attention over instances [C] //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin：ACM，2016.

[7] YANG D，WANG S，LI Z. Ensemble neural relation extraction with adaptive boosting [C] //Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence. Stockholm：ACM，2018.

[8] YOSHIDA M，IKEDA M，ONO S，et al. Person name disambiguation by bootstrapping [C] //Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval. Geneva：ACM，2010.

[9] LOUPPE G，AI-NATSHEH H，SUSIK M，et al. Ethnicity sensitive author disambiguation using semi-supervised learning [C] // International Conference on Knowledge Engineering and the Semantic Web. Prague：Springer，2016.

[10] ZHANG Y，ZHANG F，YAO P，et al. Name Disambiguation in AMiner：Clustering，Maintenance，and Human in the Loop [C] // Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining，London：ACM，2018.

[11] FAN X M，WANG J Y，LV B，et al. GHOST：an effective graph-based framework for name distinction [C] //Proceedings

of the 17th ACM Conference on Information and Knowledge Management. Napa Valley: ACM, 2008.

[12] TANG J, FONG A C M, WANG B, et al. A unified probabilistic framework for name disambiguation in digital library [J]. IEEE Transactions on Knowledge & Data Engineering, 2012, 24 (6): 975-987.

[13] ZHANG B C, HASAN M. Name Disambiguation in Anonymized Graphs using Network Embedding [C] //Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore: ACM, 2017.

[14] YIN H, BENSON A R, LESKOVEC J, et al. Local higher-order graph clustering [C] //Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM, 2017.

[15] EPASTO A, LATTANZI S, LEME R P. Ego-Splitting Framework: From Non-Overlapping to Overlapping Clusters [C] //Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Halifax: ACM, 2017.

[16] FOLINO F, PIZZUTI C. An evolutionary multiobjective approach for community discovery in dynamic networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26 (8): 1838-1852.

[17] MA X, DONG D. Evolutionary nonnegative matrix factorization algorithms for community detection in dynamic networks [J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29 (5): 1045-1058.

[18] NIU X, SI W, WU C Q. A label-based evolutionary computing approach to dynamic community detection [J]. Computer Communications, 2017, 108 (8): 110-122.

[19] WU B, ZHANG C, GUO Q. A Parallel Network Community Detection Algorithm Based on Distance Dynamics [C] // Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Sydney: ACM, 2017.

[20] ZHANG C, ZHANG Y, WU B. A Parallel Community Detection Algorithm Based on Incremental Clustering in Dynamic Network [C] //2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining. Barcelona: IEEE, 2018.

[21] 尹友明. Java语言与Java技术概述 [J]. 中国新技术新产品, 2011 (6): 99.

[22] 李盛庆, 蔡国永. 复杂网络领域科研合著网络演化及知识传播特点研究 [J]. 现代图书情报技术, 2013 (5): 64-72.

[23] 常欢, 吕瑞花, 张佳静. 学术谱系内合作网络研究——以刘东生为核心的第四纪学术谱系为例 [J]. 情报理论与实践, 2016, 39 (4): 14-19.

[24] 胡化凯, 丁兆君, 陈崇斌, 等. 当代中国物理学家学术谱系 [M]. 上海: 上海交通大学出版社, 2016: 3.

[25] WANG W, LIU J Y, XIA F, et al. Shifu: Deep Learning Based Advisor-Advisee Relationship Mining in Scholarly Big Data [C] // Proceedings of the 26th International Conference on World Wide Web Companion. Perth: 2017.

[26] BLONDE V D, GUILLAUME J L, LAMBIOTTE R, et al. Fast unfolding of communities in large networks [J]. Journal of Statistical Mechanics: Theory and Experiment, 2008 (10): P10008.

[27] SINHA A, SHEN Z, SONG Y, et al. An Overview of Microsoft Academic Service (MAS) and Applications [C] //Proceedings of the 24th International Conference on World Wide Web. Florence: ACM, 2015.

作者简介

彭程程，女，1996年生，硕士研究生，研究方向：数据挖掘、社会网络分析，E-mail：chchpeng1997@163.com。
吴斌，男，1969年生，博士，教授，博士生导师，研究方向：数据挖掘、复杂网络。

Research of Scholar Profile System: Smart Campus

PENG ChengCheng WU Bin
(Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: With the arrival of the era of big data, the academic data shows an exponential growth trend. At the same time, as a modeling method of user, user profile has been widely used in various fields recently. By analyzing and mining academic data related to scholars, we can construct a full-scale and accurate profile for scholar, which plays an important role in researching scholars' academic behavior. This paper introduces not only the scholar profile system smart campus, but also its key methods and functional features. And we compare it with other mainstream scholar profile systems. The result shows that the system has certain advantages and characteristics such as the scholars' academic pedigree and relative research.

Keywords: Scholar Profile; User Profile; Scholar Genealogy; Six Degree Search

（收稿日期：2019-01-18 ）